Advancing Biomedical Named Entity Recognition with Multivariate Feature Selection and Semantically Motivated Features

by

James Robert Leaman Jr.

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2012 by the
Graduate Supervisory Committee:

Graciela Gonzalez, Co-Chair
Chitta Baral, Co-Chair
Kevin Bretonnel Cohen
Huan Liu
Jieping Ye

ARIZONA STATE UNIVERSITY

May 2013

ABSTRACT

Automating aspects of biocuration through biomedical information extraction could significantly impact biomedical research by enabling greater biocuration throughput and improving the feasibility of a wider scope. An important step in biomedical information extraction systems is named entity recognition (NER), where mentions of entities such as proteins and diseases are located within natural-language text and their semantic type is determined. This step is critical for later tasks in an information extraction pipeline, including normalization and relationship extraction.

BANNER is a benchmark biomedical NER system using linear-chain conditional random fields and the rich feature set approach. A case study with BANNER locating genes and proteins in biomedical literature is described. The first corpus for disease NER adequate for use as training data is introduced, and employed in a case study of disease NER. The first corpus locating adverse drug reactions (ADRs) in user posts to a health-related social website is also described, and a system to locate and identify ADRs in social media text is created and evaluated.

The rich feature set approach to creating NER feature sets is argued to be subject to diminishing returns, implying that additional improvements may require more sophisticated methods for creating the feature set. This motivates the first application of multivariate feature selection with filters and false discovery rate analysis to biomedical NER, resulting in a feature set at least 3 orders of magnitude smaller than the set created by the rich feature set approach. Finally, two novel approaches to NER by modeling the semantics of token sequences are introduced. The first method focuses on the

sequence content by using language models to determine whether a sequence resembles entries in a lexicon of entity names or text from an unlabeled corpus more closely. The second method models the distributional semantics of token sequences, determining the similarity between a potential mention and the token sequences from the training data by analyzing the contexts where each sequence appears in a large unlabeled corpus. The second method is shown to improve the performance of BANNER on multiple data sets.

DEDICATION

For Chalice, who sacrificed a great deal so I could complete this dissertation, and for Ethan, Hannah and Jacob, who give my life meaning.

ACKNOWLEDGEMENTS

I owe thanks to many people for their assistance while preparing this dissertation. First, I am grateful to all members of my committee - Graciela Gonzalez, Chitta Baral, Kevin Bretonnel Cohen, Huan Liu, and Jieping Ye for supervising this dissertation. I am especially grateful to Chitta Baral and Huan Liu for helping me get started in computer science, and to Kevin Bretonnel Cohen for significant guidance and encouragement. Most of all, I am grateful to Graciela Gonzalez for her constant support over many years, for her guidance in helping me find the right direction for my research, and for her patience with my learning process. Her mentorship and the many chances she provided for me to stretch myself have been invaluable.

I owe a great deal to my collaborators: Jörg Hakenberg, Luis Tari, Laura Wojtulewicz, Ryan Sullivan, Christopher Miller, Siddhartha Jonnalagadda, Annie Skariah, Skatje Myers, and Jian Yang. I am also grateful to other lab members for helpful conversations and feedback on my work, particularly Azadeh Nikfarjam, Ehsan Emadzadeh, and Robert Yao. I appreciate the many colleagues who have provided helpful advice, useful technical discussions, and encouragement, including Lynette Hirschman, Karin Verspoor, Roman Klinger, Matthew Scotch, Violet Syrotiuk, and Rezarta Islamaj Doğan.

I am indebted to the many researchers upon whose work this dissertation builds, in particular those who have provided software implementations and datasets. I am also immensely grateful to the many researchers worldwide who have made use of my research and software - particularly BANNER - in their own work.

I am particularly grateful to Zhiyong Lu, NCBI, and the National Library of Medicine for providing me with an internship and the necessary resources to support my completion. I acknowledge the support of Science Foundation Arizona (grant CAA 0277-08), the Arizona Alzheimers Disease Data Management Core (under NIH Grant NIA P30 AG-19610), and the Arizona Alzheimers Consortium.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Like many modern sciences, the primary constraint in advancing the biological sciences is moving away from gathering data to evaluate hypotheses and instead towards the data interpretation and theory creation necessary to make sense of large amounts of data. This trend has driven an increasing recognition of the importance of biocuration - the field that organizes the results of biomedical research and makes them available [12, 56]. While this increased recognition has driven an increase in the rate of biocuration, research continues at a much faster rate than biocuration can handle [6]. As a result, the potential of existing research to enable further discoveries is not being fully realized.

One possibility for increasing the rate of biocuration is with natural language processing (NLP) techniques [1, 55, 92]. Biomedical information extraction is a sub-field of biomedical NLP that seeks to locate, categorize, and extract information from various biomedical texts, including scientific articles, in support of tasks such as biocuration, and patient records, for tasks such as clinical decision support [30].

While increasing the rate of biocuration is an important goal, another significant application of biomedical natural language processing techniques is enabling the extraction of information that would otherwise not be available. One example is mining text authored by patients for information relating to their health. The large volume of text in most social media, where such text can be found, implies the necessity of automated techniques.

1.1 Overview of Information Extraction

Biomedical information extraction systems are typically designed as pipelines, with each module in the pipeline performing a specific processing task. Some tasks are relatively straightforward and typically handled by deterministic techniques. These include sentence segmentation and tokenization (breaking text into individual word-like units). Other tasks involve more nuanced decision making and frequently utilize machine learning, typically supervised classification. In a typical biomedical information extraction pipeline, the first task encountered that may require advanced techniques is named entity recognition (NER). NER is the task of locating mentions of entities in natural language text, specifying both the span - start and end position - and semantic type [74].

While many methods have been used for biomedical NER, the state-of-the-art generally involves tokenization as a preprocessing step, followed by labeling with a supervised sequence classification model using a rich feature set [69, 100]. Rich feature sets describe a wide variety of different aspects of each token, including prefixes and suffixes, word stems or lemmas, part of speech, and so on [100]. These features are typically binary-valued; an example would be whether the current token ends in "-ase." However, in the rich feature set approach the NER system developer typically does not create individual features. Instead, developers create feature extraction templates, such as "the last three characters of each token." These are then instantiated into binary features using the actual values seen in the training data.

1.2 Problem Background

The rich feature set approach has been successfully applied to find many entity types, including genes, proteins, RNA, cell lines and cell type [100], genomic variations [79], drug names [64], and diseases [34, 70]. A critical advantage of the rich feature set approach is the ability to adapt the feature set to the dataset used. The rich feature approach introduces several difficulties, however, that result in diminishing returns as the feature set is developed. This chapter argues that the feature extraction methods used to create rich feature sets for biomedical named entity recognition have become a constraint for improved performance. This discussion is motivated through an analysis of several qualities that contribute to system performance and how the rich feature set approach affects each.

A system is said to generalize well if it performs well on previously unseen data [2]. The distributions of the frequency that specific words appear in a text approximates a power law, a result known as Zipf's law [77, 128]. According to Zipf's law, the frequency a word appears in a large segment of text is inversely proportional to the rank of its frequency. This implies a "long tail" effect where a few words appear frequently, some words have medium frequency, but most words are rare. Many of the tokens in the unseen text have therefore not been seen in the training data. Since the features are extracted from the training data, the model therefore only contain a few features usable for inference on these tokens, which increases overfitting and reduces the generalization of the model.

A system is stable if small variations in the training data produce models with low levels of disagreement when applied to the same text [101].

Zipf's law also implies that most words present in the training data appear only a few times, causing most of the features to have a highly skewed ratio between the number of times the feature is active and inactive. This imbalance causes the correlations of the features with the class to be sensitive to small perturbations in the training data, thereby reducing model stability.

A system is robust if the performance declines gracefully with the introduction of noise [39]. For NER, the training data is manually created and is difficult to keep consistent [1], making it a significant source of noise. However, because categorical variables are represented in the feature set as many boolean features with only one feature with the value true - the one-hot representation - inactive features are relatively uninformative for classification in rich feature sets, and only a very small percentage of the features in the set will be active for any given token. The model is therefore forced to classify each token using whatever features happen to be active, some of which are not stable. This reduces robustness by increasing the model sensitivity to any inconsistencies in the training data.

A system has higher learning efficiency if it achieves higher performance given the same amount of training data [124]. High learning efficiency is particularly important when only a small amount of data is available, but is always a concern because of the high expense of annotating new corpora to obtain training data. Training data is required to estimate the relevance of each feature, so that learning efficiency is reduced as the number of features increases. Unfortunately, rich feature sets typically reach extremely high dimensionality since the features are derived from the words in the training set, and most of the words are rare. It is common for state-of-the-art biomedical NER systems to use hundreds of thousands or

even millions of features. In such a large feature set, many of the correlations between the features and the class are likely due to chance [90], a problem known in statistics as multiple comparisons [60].

The portability of a system concerns the amount of new development required to obtain quality results in a different corpus or domain. Porting an existing biomedical NER system to recognize mentions of a new entity type usually requires analysis to identify new feature templates. Entity mentions frequently exhibit a meaning that is not entirely compositional, similar to collocations [77]. In other words, the meaning of the sequence of tokens that comprise the name - in this case, whether the sequence refers to an entity of a specified type - is richer than the union of the meanings of each constituent token. Unfortunately, the sequence classification models typically in use in biomedical named entity recognition can only model sequence features that can be decomposed by the Markov assumption [98]. Since this extra-compositional meaning cannot be modeled directly, the model must compensate by relying more on indirect clues such as features from manually engineered templates, thus decreasing the portability.

1.3 Overall Goal

The primary goal of this dissertation is to improve biomedical named entity recognition (NER). This improvement is accomplished through two complementary approaches, based on the premise that the rich feature representation currently used in state-of-the art biomedical NER is subject to diminishing returns. The first approach uses advanced feature selection techniques to determine the relevance of each feature extracted and remove features that prove uninformative. The second approach introduces

sophisticated new features to model the meaning of the token sequences more closely than the existing binary features.

1.4 Contributions

There are several innovate aspects of this work. Distributional semantics has been shown to be useful for improving NER [61, 115]. However all techniques used to date are based on single tokens, even though entity names exhibit non-compositional meanings. This motivates the first distributional semantics technique for modeling the semantics of token sequences in the context of NER, rather than individual tokens.

Biomedical NER systems have been shown to not always benefit when features derived from a list of entity names (a dictionary) were used [100]. Processing the name list to only contain highly indicative tokens was able to show an improvement [50], an approach that has been both automated and strengthened theoretically by incorporating language modeling of both the entity names and general biomedical text.

Previous work applying feature selection to biomedical NER used the χ^2 test, a standard statistical hypothesis test, and information gain, an information-theoretic criterion [62]. Since NER features are highly imbalanced, however, χ^2 tests are not appropriate for determining feature significance [35, 91]. While much of the literature on feature selection assumes that feature redundancy is detrimental, more recent work shows through theoretical and empirical analysis that this is not accurate [13]. Rich feature sets exhibit a very high degree of redundancy since they are generated by applying many interdependent extraction templates to the training data. The application of feature selection therefore employs a filter based feature

selection algorithm which considers feature redundancy, resulting in the first application of joint mutual information to biomedical NER.

The false discovery rate (FDR) - roughly the percentage of the features accepted as relevant which are actually irrelevant - has been shown to be a useful criterion for determining a stopping threshold for feature selection [45]. The first FDR analysis of a feature selection algorithm in the context of biomedical NER is performed.

Chapter 2

FUNDAMENTALS

Biomedical information extraction systems typically use the pipeline architecture, with biomedical NER used as a building block step for higher level information extraction tasks including entity identification and relation extraction. NER is therefore a critical task for biomedical information extraction from biomedical texts and for most forms of natural language processing. While there has been significant work to solve NER, performance for NER systems in the biomedical domain is still significantly lower than human performance. The primary goal is to identify the remaining challenges and propose new work to resolve them.

2.1 Information Extraction

Natural language processing of biomedical articles has received significant attention. Much of the work has been driven by academic challenges such as BioCreative and the BioNLP shared tasks [3, 72, 104]. Many of these tasks have concentrated on applications of biomedical information extraction, particularly entity identification, protein-protein interaction extraction and event detection.

2.2 Named Entity Recognition

In named entity recognition (NER), the task is to locate the entities referenced in natural language text, and determine the semantic type of each. Each reference to an entity is called a mention. An NER task in the newswire domain might require the location of mentions referring to people,

places and organizations. NER in the biomedical domain typically involves semantic types such as proteins, genes, diseases, organisms, and drugs. Biomedical NER is a particularly difficult problem for several reasons:

1. There are many semantic types of interest to researchers; the UMLS Metathesaurus, for example, contains 135 different semantic types [87].

2. There are many names used in the literature. For example, there are millions of gene names in actual usage [104]. Many entities have several names in actual usage. This is partly because authors often prefer to use a name of their own invention rather than an official or standardized name for the entities their writing refers to [36].

3. Names are often ambiguous. This is especially a problem with acronyms [74]. Thus the correct semantic type often must be inferred from context even if the name itself is recognized. "HD" for example, could refer to either Huntington disease or the gene whose mutated form causes it [70].

4. Many semantic types are easily confused with each other due to similar vocabulary or context. For example, trained human annotators cannot always distinguish between genes and proteins [110]. In fact, whether an entity should be considered a separate type or not depends on the purpose of the task.

2.3 Techniques for Named Entity Recognition

Methods for NER fall into three primary categories [74]; each category is surveyed in this section. The first method considered is the so-called dictionary approach, where a list of entity names is used to locate the

entities in biomedical natural language text. Typically some processing is performed, such as case normalization or handling of variant terms or transformations that are specific to the entity type. For example, a lower case "h" at the beginning of a gene name indicates that the mention refers to the gene as found in humans. Dictionary methods have the advantage of immediately providing a potential identification of the entity being referenced by the mention. On the other hand, they have the disadvantage of requiring a comprehensive name list that is not available for all entity types of interest. It can also be difficult to create the set of string transformations that represent valid variations of the names. Despite these difficulties, dictionary methods for NER enjoy widespread support.

A second common method is the rule based approach, where a set of patterns - frequently based on regular expressions - is applied to locate entities of the specified type. This approach has the advantage of not requiring a comprehensive listing of entity names. It has the additional advantage of providing an explanation for the decisions it makes; the rule that triggered the decision can be used to show the user why the decision was made. The primary disadvantage of this approach is the significant investment required to create the rules. In addition, these rules have been found to not generalize well: performance drops significantly when applied to text with slightly different characteristics.

A third approach is the use of machine learning in the form of supervised classification. Early approaches used instance classifiers such as naïve bayes or support vector machines to predict the label for a given token [74]. Later work introduced sequence classification models including hidden Markov models (HMMs) and maximum entropy Markov models (MEMMs),

which learn a mapping from an input sequence to a sequence of output labels [11, 23]. Unlike HMMs, MEMMs are discriminative models and therefore do not assume independence between the features. This allows NER system developers to utilize virtually any feature they believe may be useful. However, this strong advantage of discriminative models is partially negated by the dependence on the training data introduced by limiting the feature set to the values observed there.

Many systems employ some form of hybrid between the three techniques. As an example, a recent work used a combination of linguistic rules and a large number of dictionaries to assign a broad semantic class to a large number of potentially overlapping semantic classes [20].

2.4 Machine Learning Methods for NER

NER systems employing classification must choose a set of labels to differentiate between mentions of different types and non-mention text [102]. The simplest method uses one label ("O") for non-mention tokens, and another label (e.g. "I-Protein" or "I-Drug") for each entity type in the training data, however this label set cannot differentiate between adjacent entities of the same type. While this condition is actually rare [37], many systems report a performance increase by employing a different label ("B") for the first token in a mention.

2.4.1 Rich Feature Sets

All machine learning systems receive their input about a learning task via the set of features used. These features are critically important, as the learning algorithm will be blind to any information they do not provide.

Each feature must be encoded into the representation required by the machine learning component. For most machine learning systems, features are encoded as a vector of numeric values which, for convenience, are often indexed by labels representing the meaning of the feature. For example, a label could be token='gene', indicating that the token is the word "gene", and the value would be binary, with 0 representing false and 1 representing true. For NER, each input text - typically a single sentence - is represented by a sequence of feature vectors.

Current state-of-the-art systems for biomedical NER typically utilize a rich feature set with a size in the hundreds of thousands [69, 100]. This often results in more features than points of training data, a condition known as a wide dataset. Most biomedical NER systems therefore use regularization to control overfitting. Regularization introduces a penalty on large parameters in an attempt to keep the weight of a few parameters from overwhelming the remainder. The most common form of regularization used in biomedical NER systems appears to be $\mathcal{L}_2$ regularization, which drives the parameters for irrelevant features towards zero asymptotically. Other forms of regularization are also possible, notably $\mathcal{L}_1$ regularization, which results in a sparse solution since the parameters for useless features are driven to exactly zero [68, 89].

For example, the system developer may create a template that instantiates a feature from each of the suffixes of length 3 seen in the training data. Some of these features will be informative - for example, tokens that end in the suffix "-ase" are frequently names of enzymes, a kind of protein - while many features will be marginal or irrelevant. The resulting features are typically encoded using the one-hot representation, where only one of the features generated by a template, such as "suffix=ase", will be active (have

the value true) and all others will be inactive (have the value false). Achieving the highest possible performance on a new dataset requires significant new development to engineer new feature templates. One example from a work to recognize chemical names in text is the inclusion of features representing the single characters before and after the current token [64].

2.4.2 Conditional Random Fields

Much of the recent work in biomedical NER has centered on a discriminative sequence classifier, linear-chain conditional random fields [67, 79]. Unlike MEMMs, conditional random fields (CRFs) are normalized per sequence rather than per tag, avoiding the so-called "label bias". While conditional random fields can, in general, take the form of an arbitrary graph, the form most often used in natural language processing is linear-chain, where the nodes are arranged in a sequences and only connected to adjacent nodes.

Following the notation in [63], the equations for conditional random fields are:

$$p_{\overrightarrow{\lambda}}(\overrightarrow{y}|\overrightarrow{x}) = \frac{1}{Z_{\overrightarrow{\lambda}}(\overrightarrow{x})} \cdot exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y_{j-1}, y_j, \overrightarrow{x}, j)\right)$$

$$Z_{\overrightarrow{\lambda}}(\overrightarrow{x}) = \sum_{\overrightarrow{y}\in\mathcal{Y}(m)} exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y_{j-1}, y_j, \overrightarrow{x}, j)\right)$$

Where:

- $\overrightarrow{x}$ is the input sequence
- $\overrightarrow{y}$ is the sequence of output labels
- $n = |\overrightarrow{x}| = |\overrightarrow{y}|$
- $\overrightarrow{\lambda}$ are the feature weights

- f are the feature functions
- $m = |\overrightarrow{\lambda}| = |f|$
- $\mathcal{Y}$ is the set of possible labelings

The time complexity for training and inference for linear-chain conditional random fields are as follows:

- $O(tks^2n)$, for training
- $O(s^2n)$, for inference

Where:

- t is the number of training instances
- s is the number of states
- k is the number of training iterations performed
- n is the length of the instance

2.5 NER System Evaluation

Biomedical NER systems are typically evaluated in terms of precision (p) and recall (r), which is then frequently summarized in the F_1 measure (f). These are defined as follows:

$$p = \frac{tp}{tp + fp}, \quad r = \frac{tp}{tp + fn}, \quad F_1 = \frac{2pr}{p + r}$$

Where:

- tp is the number of true positives

- fp is the number of false positives
- fn is the number of false negatives

These definitions do not specify, however, what constitutes a true positive, false positive and false negative. In NER these definitions are not always straightforward, since it is arguably better for a system to find most of the span of a mention than to not mark it at all or perhaps to mark a mention with a slightly incorrect semantic type rather than to miss it completely. This leads to many possibilities in how to consider a match correct, however the most common evaluation measure is what is known as exact match [41]. Exact match requires that the left boundary, the right boundary and the semantic type all match exactly for a true positive to be counted. Any mention returned by the system which is not a true positive is counted as a false positive, and any mention required by the evaluation data which does not have a corresponding true positive is considered a false negative.

2.6 Corpora for NER Training and Evaluation

Modeling biomedical NER as a supervised learning problem implies that training data will be needed in addition to data for evaluation. Several biomedical NER corpora are available for training and evaluation. These corpora are annotated for different semantic types, and represent different degrees of size and quality.

The BioCreative 2 Gene Mention corpus contains sentences from biomedical abstracts annotated with genes and proteins as a single semantic type. This corpus also contains alternate annotations; the scoring for this

corpus is therefore modified so that any one of the alternate annotations is considered a true positive. The NCBI Disease Corpus contains complete biomedical abstracts annotated for disease mentions. This corpus is derived from previous work for disease mentions by the author [70]. The BioCreative 2 Gene Mention and NCBI Disease corpora form the core of the evaluation in this work. There are many other corpora, however. One interesting recent corpus is the CALBC corpus, which was created by harmonizing the annotations of multiple automated systems [48]. Because this corpus was not created by human annotators, it is called a silver standard corpus. The advantage of a silver standard corpus, however, is the feasibility of providing much more data than may be provided by human annotators.

Chapter 3

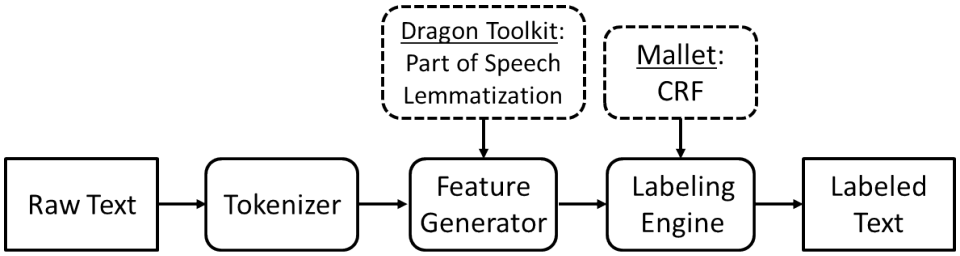
CASE STUDY: GENES AND PROTEINS

This chapter describes a case study in biomedical named entity recognition, locating genes, proteins, and diseases. This work resulted in the creation of BANNER, a trainable biomedical NER system based on conditional random fields a rich feature set.

3.1 Background

BANNER is an open source biomedical named entity recognition system implemented in Java, serving as an executable survey of advances [69]. BANNER based on conditional random fields using the rich feature set approach. BANNER implements a wide range of orthographic, morphological, and shallow syntax features, including the part of speech, lemma, n-grams, prefixes, and suffixes. A primary design constraint for BANNER is configurability: BANNER is intended to enable experimental evaluation of a variety of different configurations, including the label model and the order. The initial version of BANNER did not include a feature based on lists of entity names, but did include two forms of postprocessing. The first form of postprocessing detects when only one of a pair of parentheses was tagged in the output. The second form of postprocessing is detection of a long form, short form pair, such as "antilymphocyte globulin (ALG)" [99].

Figure 3.1: The architecture of BANNER.



3.2 Methods

BANNER is designed as a processing pipeline. Input text is first broken into sentences, and is then tokenized. BANNER uses a tokenization strategy that is both straightforward and highly consistent. Tokens are broken at all white space and at punctuation. The tokens returned therefore consist of contiguous alphanumeric sequences or a single punctuation mark.

A series of experiments is performed to determine the highest-performing configuration of BANNER for the BioCreative 2 Gene Mention data set. The configuration using a 2nd order CRF, the IOB label model, using parenthesis post-processing, and not splitting tokens at letter/digit boundaries were the best performing configuration elements. A series of experiments to manually select feature templates are performed, where it was found that adding part of speech tags, lemmas, and numeric normalization all improved performance.

The BANNER architecture is a 3-stage pipeline, illustrated in Figure 3.1. Input is taken one sentence at a time and separated into tokens, contiguous units of meaningful text roughly analogous to words. The stream of tokens is converted to features, each of which is a name/value pair for use

by the machine learning algorithm. The set of features encapsulates all of the information about the token the system believes is relevant to whether or not it belongs to a mention. The stream of features is then labeled so that each token is given exactly one label, which is then output.

The tokenization of biomedical text is not trivial and affects what can be considered a mention since generally only whole tokens are labeled in the output [125]. Unfortunately, tokenization details are often not provided in the biomedical named entity recognition literature. BANNER uses a simple tokenization which breaks tokens into either a contiguous block of letters and/or digits or a single punctuation mark. For example, the string "Bub2p-dependent" is split into 3 tokens: "Bub2p", "-", and "dependent". While this simple tokenization generates a greater number of tokens than a more compact representation would, it has the advantage of being highly consistent.

BANNER uses the CRF implementation of the MALLET toolkit [78] for both feature generation and labeling using a second order CRF. The set of machine learning features used primarily consist of orthographic, morphological and shallow syntax features. While many systems use some form of stemming, BANNER instead employs lemmatization [119], which is similar in purpose except that words are converted into their base form instead of simply removing the suffix.

Another notable feature is the numeric normalization feature [112], which replaces the digits in each token with a representative digit (e.g. "0"). Numeric normalization is useful since entity names often occur in series, such as the gene names Freac1, Freac2, etc. The numeric-normalized value for all these names is Freac0, so that forms not seen in the training data have the

same representation as forms which are seen. The entire set of features is used in conjunction with a token window of 2 to provide context, that is, the features for each token include the features for the previous two tokens and the following two tokens.

There are features discussed in the literature which are not implemented in BANNER, particularly semantic features such as a match to a dictionary of names and deep syntactic features, such as information derived from a full parse of each sentence. Semantic features generally have a positive impact on overall performance [125] but often have a deleterious effect on recognizing entities not in the dictionary [100, 127]. Moreover, employing a dictionary reduces the flexibility of the system to be adapted to other entity types, since comparable performance will only be achieved after the creation of a comparable dictionary. While such application-specific performance increases are not the purpose of a system such as BANNER, this is an excellent example of an adaptation which researchers may easily perform to improve BANNER's performance for a specific domain.

Deep syntactic features are derived from a full parse of the sentence, which is a noisy and resource-intensive operation with no guarantee that the extra information derived will outweigh the additional errors generated [74]. The use of deep syntactic features in biomedical named entity recognition systems is not currently common, though they have been used successfully. One example is the system submitted by Vlachos to BioCreative 2 [119], where features derived from a full syntactic parse boosted the overall F-score by 0.51.

There are, however, two types of general post-processing which have good support in the literature and are sufficiently generic to be applicable to

any biomedical text. The first of these is detecting when matching parentheses, brackets or double quotation marks receive different labels [32]. Since these punctuation marks are always paired, detecting this situation is useful because it clearly demonstrates that the labeling engine has made a mistake. BANNER implements this form of processing by dropping any mention which contains mismatched parentheses, brackets or double quotation marks. The second type of generally-applicable post-processing is called abbreviation resolution [127]. Authors of biomedical articles often introduce an abbreviation for an entity by using a format similar to "antilymphocyte globulin (ALG)" or "ALG (antilymphocyte globulin)". This format can be detected with a high degree of accuracy by a simple algorithm [99], which then triggers additional processing to ensure that both mentions are recognized.

3.3 Comparison

BANNER was evaluated with respect to the training corpus for the BioCreative 2 GM task, which contains 15,000 sentences from MEDLINE abstracts and mentions over 18,000 entities. The evaluation was performed by comparing the system output to the human-annotated corpus in terms of the precision (p), recall (r) and their harmonic mean, the F-measure (F). These are based on the number of true positives (TP), false positives (FP) and false negative (FN) returned by the system:

The entities in the BioCreative 2 GM corpus are annotated at the individual character level, and approximately 56% of the mentions have at least one alternate mention annotated, and mentions are considered a true positive if they exactly match either the main annotation or any of the

alternates. The evaluation of BANNER was performed using 5x2 cross-validation, which Dietterich shows to be more powerful than the more common 10-fold cross validation [31]. Differences in the performance reported are therefore more likely to be due to a real difference in the performance of the two systems rather than a chance favorable splitting of the data.

The initial implementation of BANNER included only a naïve tokenization which always split tokens at letter/digit boundaries and employed a 1st-order CRF. This implementation was improved by changing the tokenization to not split tokens at the letter/digit boundaries, changing the CRF order to 2, implementing parenthesis post-processing and adding lemmatization, part-of-speech and numeric normalization features. Note that both the initial and final implementations employed the IOB label model. Table 3.1 presents evaluation results for the initial and final implementations, as well as several system variants created by removing a single improvement from the final implementation.

The only system variant which had similar overall performance was the IO model, due to an increase in recall. This setting was not retained in the final implementation, however, due to the fact that the IO model cannot distinguish between adjacent entities. All other modifications result in decreased overall performance, demonstrating that each of the improvements employed in the final implementation contributes positively to the overall performance.

The performance of BANNER was compared against the existing freely-available systems in use, namely ABNER [100]. The evaluations are performed using 5x2 cross validation using the BioCreative 2 GM task training corpus, and reported in Table 3.2.

BANNER System Variant	Precision	Recall	F-measure
Initial implementation	0.8239	0.7621	0.7918
Final implementation	0.8509	0.7906	0.8196
With IO model instead of IOB	0.8471	0.7940	0.8196
Without numeric normalization	0.8456	0.7909	0.8174
With IOBEW model instead of IOB	0.8546	0.7815	0.8164
Without parenthesis post-processing	0.8509	0.7906	0.8196
Using 1st order CRF instead of 2nd order	0.8449	0.7872	0.8150
With splitting tokens between letters and digits	0.8454	0.7835	0.8133
Without lemmatization	0.8444	0.7800	0.8109
Without part-of-speech tagging	0.8402	0.7783	0.8081

Table 3.1: Results of evaluating the initial version of BANNER, the final version, and several system variants created by removing a single improvement from the final implementation.

System	Precision	Recall	F-measure
BANNER	0.8509	0.7906	0.8196
ABNER	0.8312	0.7394	0.7830

Table 3.2: Results of comparing BANNER against existing freely-available software, using 5x2 cross-validation on the BioCreative 2 GM task training corpus.

To demonstrate portability, another experiment is performed using 5x2 cross validation on the disease mentions of the BioText disease-treatment corpus [94]. These results are reported in table 3.3. The relatively low performance of all three systems on the BioText corpus is likely due to the small size (3655 sentences) and the fact that no alternate mentions are provided.

Like BANNER, ABNER is also based on conditional random fields; however it uses a 1st-order model and employs a feature set which lacks part-of-speech, lemmatization and numeric normalization features. In

System	Precision	Recall	F-measure
BANNER	0.6889	0.4555	0.5484
ABNER	0.6608	0.4486	0.5344

Table 3.3: Results of comparing BANNER against existing freely-available software, using 5x2 cross-validation on the disease mentions from the BioText corpus.

System or Author	Rank at BioCreative 2	Precision	Recall	F-measure
Ando	1	0.8848	0.8597	0.8721
Vlachos	9	0.8628	0.7966	0.8284
BANNER	-	0.8718	0.8278	0.8492
Baumgartner et. al.	11 (median)	0.8554	0.7683	0.8095
NERBio	13	0.9267	0.6891	0.7905

Table 3.4: Comparison of BANNER to selected BioCreative 2 systems [104].

addition, it does not employ any form of post-processing, though it does use the same IOB label model. ABNER employs a more sophisticated tokenization than BANNER, however this tokenization is incorrect for 5.3% of the mentions in the BioCreative 2 GM task training corpus.

The large number of systems (21) which participated in the BioCreative 2 GM task in October of 2006 provides a good basis for comparing BANNER to the state of the art in biomedical named entity recognition. These results are reported in Table 3.4.

The performance of the BANNER named entity recognition system was later increased by 1.4% f-measure on the BioCreative 2 Gene Mention set by adding a list of single tokens highly indicative of a gene name being present [50]. This list was derived by extracting all of the single tokens from the gene and protein names in EntrezGene, UniProt, HUGO and the BioCreative 2 Gene Normalization training set. Tokens more likely to appear

outside of an entity mention than inside of an entity mention in the training data for the BioCreative 2 Gene Mention training data were then removed. The list is used as a binary feature, with tokens on the list (case-insensitive) given the value 1, and all others 0.

3.4 Conclusion

BANNER, an executable survey of advances in named entity recognition, has been shown to achieve significantly better performance than existing open-source systems. This is accomplished using features and techniques which are well-supported in the more recent literature. In addition to confirming the value of these techniques and indicating that the field of biomedical named entity recognition is making progress, this work demonstrates that there are sufficient known techniques in the field to achieve good results using known techniques.

This system is anticipated to be valuable to the biomedical NER community both by providing a benchmark level of performance for comparison and also by providing a platform upon which more advanced techniques can be built. It is also anticipated that this work will be immediately useful for information extraction experiments, possibly by including minimal extensions such as a dictionary of names of types of entities to be found.

Chapter 4

CASE STUDY: DISEASES

Many interesting questions involving text mining require the location and identification of diseases mentioned in biomedical text. Examples include the extraction of associations between genes (or gene mutations) and diseases or between drugs and diseases. Named entity recognition is the problem of locating mentions in a text and tagging them with their semantic type, in this case "disease." Normalization is the process of determining which specific entity the mention refers to, often by returning a unique identifier associated with the concept. For example, the unique identifier (CUI) for "myocardial infarction" in the UMLS Metathesaurus is C0027051 [87]. Note that all mentions in any given text are needed for automatic processing, along with their approximate locations, to determine their associations or relationships.

It has been recognized previously that biomedical entities such as genes and proteins suffer from several problems when locating and identifying them in biomedical text [104]. These problems include:

- The large number of names in use
- Multiple names used to refer to the same entity
- A single name used to refer to more than one entity of the same semantic type
- Similar or identical names used to refer to entities of differing semantic types
- Complex syntactic structures used to refer to multiple related entities

These issues may also be relevant for the recognition and identification of disease mentions, and that there may be other problems which are more or less unique to this semantic type. This chapter explores the extent to which this is true.

4.1 Related Work

There has been a significant amount of work related to locating and identifying diseases in various kinds of biomedical text. The work closest to the present effort is the corpus of 597 sentences from MEDLINE annotated with disease concepts from the UMLS Metathesaurus which was created by Jimeno et al. (2008), and is freely available online. The authors utilize this corpus to evaluate a dictionary approach, a statistical approach and an application of MetaMap [4] to identify diseases within the corpus. The authors found that dictionary lookup results in an f-measure of 0.593, their statistical method an f-measure of 0.280 and utilizing MetaMap results in an f-measure of 0.307. However, the corpus lacks annotation of the disease mention locations, which are needed for NER.

The corpus by Jimeno et al. was also used by Névéol et al. to improve the application of MetaMap and also evaluate an additional technique called the priority model [88, 109]. The authors break the corpus into a training set with 276 sentences and a test set with 275 sentences. The authors report the performance of the priority model at 0.80 precision, 0.74 recall and 0.77 f-measure. For the improved MetaMap method, the authors report 0.75 precision, 0.78 recall, and and 0.76 f-measure. The authors conclude that both techniques are effective, however this study is somewhat limited by the small size of the test set used. In addition, every test sentence included at

least one disease mention, which may have the unfortunate effect of reporting higher precision than would be found if the same techniques were applied to arbitrary biomedical text.

The BioText corpus contains 3,655 sentences taken from MEDLINE titles and abstracts [94]. It is annotated for location of disease and treatment mentions but not concepts, and it is freely available online. The primary goal of this corpus, however, was to explore the different types of relationships between the diseases and treatments found, so that a high degree of annotation consistency was not not required at the token level. This is likely to cause the performance of systems using it to be overly pessimistic. An assessment of BANNER, a named entity recognition system discussed further in section 4.2, reported a performance of only 0.584 f-measure when trained on this corpus [69].

The PennBioIE corpus contains 2,514 PubMed abstracts annotated for tokens, parts of speech and mention location [66], and is also available online. The corpus focuses specifically on two biomedical sub-domains, oncology and cytochrome P450 enzymes, and does not contain the disease concept annotations needed for normalization.

Chun et al. utilize the UMLS Metathesaurus to create a disease dictionary, and combine several sources to create a gene dictionary [18]. They then use these to tag a corpus of 1,362,285 MEDLINE abstracts, from which they randomly select 1,000 sentences containing both a gene mention and a disease mention for annotation by a biologist. These sentences are used to create a maximum-entropy model to filter out false positives reported by the dictionary. This technique was shown to be successful for improving the precision of the gene/disease relations found relative to the baseline of

simple co-occurrence. The highest precision reported for disease NER is 90.0%, however recall is only reported relative to the recall of the initial dictionary technique, indicating a 3.4% drop.

In addition to corpora, there are several systems which are capable of identifying diseases in free text. MetaMap has already been mentioned. Another example is Whatizit [93], which is a web service that allows a variety of text processing modules to analyze text for the information they contain. Whatizit currently offers three modules that can be used for locating disease mentions in text. Two are dictionary approaches, one of which utilizes the UMLS Metathesaurus, and the other a lexicon from healthcentral.com. The third module offers a front end to the MetaMap system.

The primary contribution of this chapter is the creation, description and release of a corpus of sentences from biomedical research articles which contains annotations for both disease mention location and identification of disease concepts. This corpus can be freely downloaded online. Moreover, the experiment has demonstrated that the corpus is large enough to allow training and evaluation of machine learning-based named entity recognition systems. The similarities and differences in the automatic processing of disease mentions as compared to other biomedical entities, particularly genes and proteins, is also noted. As far as the authors are aware, no existing corpus offers a substantially similar combination of properties.

4.2 Corpus

This section presents the methodology for creating the disease corpus. Corpus statistics and analysis are also discussed. The 597 sentences sampled by Jimeno et al. sampled from the corpus by Craven and Kumlien [25] were

selected to start. Mention location annotations were then added and corrected some concept identifiers to be most specific rather than most general. When concept annotation differs that of Jimeno et al., it is noted in the corpus. Additional sentences were then selected from the Craven corpus, for at total of 2,784 sentences. The Craven corpus contains sentences selected from MEDLINE abstracts via a query for six proteins. These were originally annotated for disease concepts from OMIM as part of an analysis of gene-disease relationships [25].

Each sentence was annotated for the location of all disease mentions, including duplicates within the same sentence. The location of the mention was taken to be the minimum span of text necessary to include all the tokens required for the most specific form of the disease. The disease mention for the phrase "insulin-dependent diabetes mellitus" was therefore taken to be the entire phrase, rather than simply "diabetes mellitus" or "diabetes." It was determined, however, that any mention of the organism or species should not be included. Thus, the mention for "human X-linked recessive disorder" would not include "human." Local abbreviations such as "Huntington disease (HD)" were annotated as two separate mentions. Each mention was also mapped - where possible - to a unique concept (CUI) in the UMLS Metathesaurus from one of the following types:

- Disease or syndrome
- Neoplastic process
- Congenital abnormality
- Acquired abnormality

- Experimental model of disease
- Injury or poisoning
- Mental or behavioral dysfunction
- Pathological function
- Sign or symptom

The type "sign or symptom" refers to entities which are not actually diseases. However new diseases are often referred to as a set of signs and symptoms until the disease receives an official name.

If multiple disease concepts were judged appropriate, the most specific concept justifiable from the text or its context was used. For example, a mention of "type 2 diabetes" would be annotated as "type II diabetes mellitus" (C0011860) rather than the less-specific "diabetes mellitus" (C0011849). It was not always possible to identify the most-specific concept corresponding to the mention from the mention itself, necessitating the incorporation of the surrounding context in the determination.

The corpus includes a notes field that indicates whether the mapping is "textual" or "intuitive." A textual mapping means that the text of the mention was sufficient to find the concept in the UMLS Methathesaurus. An intuitive mapping means that either context from the abstract or a subset of the mention terms was used to locate the associated concept. Detailed comments specify what was needed or what logic was followed to determine the annotation applied for all intuitive annotations.

Disease names embedded in entities of other types were not annotated as referring to the disease. Thus references to the "HD gene" were not taken

to be mentions of "HD," which is the typical abbreviation for Huntington disease. Coordinations such as "Duchenne and Becker muscular dystrophy" were annotated with separate, but overlapping, disease mentions, in this case "Duchenne and Becker muscular dystrophy" and "Becker muscular dystrophy." These are then mapped to the corresponding disease concepts.

Generic words such as "disease" or "syndrome" were not annotated as disease mentions on their own, though they were included as part of a mention that was otherwise valid.

4.3 Methods

This section describes the techniques used for named entity recognition of diseases. The evaluation methodology and the results are also described.

4.3.1 Dictionary Techniques

The first method tested is a lexical approach through the use of dictionary lookup. This method simply performs an exact match between the concepts in the dictionary to the terms present in the free text. The dictionary was comprised of the names listed in the UMLS Metathesaurus from the types which were used to annotate the corpus.

An advantage of this method is that it automatically performs normalization to some extent in that terms are only found if they exactly fit the concepts defined in the dictionary. Additionally, it is also computationally very fast. Disadvantages are that lexical variations of terms in the dictionary cannot be dealt with and any terms not present in the dictionary cannot be found, causing methods of this sort to suffer from low recall. Dictionary approaches also suffer from not being able to handle

ambiguous terms, causing a reduction in precision. Normalization techniques, such as stemming tokens, will increase recall, and filtering techniques, such as those utilized by Chun et al. increase precision [18]. All such techniques do increase system complexity, however. Dictionary approaches have the advantage of being straightforward to implement and are frequently used in practice. They therefore also form a useful baseline with which to compare NER systems implementing more sophisticated methods.

4.3.2 Conditional Random Field Systems

Conditional random fields is a machine learning technique which has been applied to named entity recognition by many successful systems [104]. As a supervised machine learning technique, it requires both training data and a set of features. It is typically used to model NER as a sequence label problem, where the input is a sequence of feature vectors and the output is a sequence of labels from a predefined set. The labels are used indicate whether the token is a part of a mention, and its type if so. CRFs model the dependencies between the previous labels and the current label, and the number of previous labels considered is called the order.

BANNER is a named entity recognition system based on conditional random fields [69]. BANNER was upgraded for this study to use a dictionary as an input feature. Improvements were also made to facilitate loading data from differing formats for training and testing, and the updates are available online*. The BANNER system uses a 3-stage pipeline as its method in identifying named entities. The first stage is tokenization where BANNER implements a simple method creating tokens by splitting contiguous characters/digits at white space and punctuation marks. Next, feature

generation including lemmatization (where words are transformed to their base forms), part of speech tagging, N-grams, and others. The final step is the use of the labeling engine utilizing conditional random fields as implemented in the MALLET toolkit [78]. BANNER allows the training of a 1st or 2nd order model, both of which are evaluated. BANNER can also make use of a dictionary as input. The dictionary used is comprised of the names for all of the concepts contained in the UMLS Metathesaurus under the types which were used for annotating the corpus, and is therefore identical to the one employed for the dictionary approach.

The Julie Lab Named Entity Tagger, or JNET, is another named entity recognition system based on conditional random fields [47]. JNET is intended to be generic enough to be trained to recognize any entity type relevant in biomedical text. It has a configurable feature set and while it can make use of part of speech information, this must be provided externally. JNET also does not provide the ability to use a dictionary. For this work, JNET was modified to use the same tokenization as BANNER, which has the effect of removing performance differences due to tokenization.

BANNER has been further extended with features based on a full parse of each sentence using the Link Grammar parser [85]. Link Grammar is a hybrid lexical and rule-based system for syntactic parsing that does not use statistical techniques [103]. Previous work had shown a small but consistent improvement in biomedical NER performance when syntactic parse features are added [105, 118], but had not evaluated the Link Grammar parser specifically. The work added several new feature templates, based on the part of speech for each. One pair of feature templates indicated the set of adjectives modifying a noun, and the noun being modified by each adjective.

Item	Count
Abstracts	794
Sentences	2,784
Tokens	79,951
Disease mentions (total)	3,228
Disease mentions (unique)	1,202
Disease concepts	686

Table 4.1: Size of the Arizona Disease Corpus, by several forms of measurement.

Another feature template indicates the verb for an object. These features were motivated by the desire to introduce longer-range dependencies into the feature set, and resulted in a small but consistent improvement in both precision (0.95%) and recall (0.58%), for improvement in f-measure (0.75%).

4.4 Results

This section describes the corpus and the results of the text mining study.

4.4.1 Corpus statistics

The size of the corpus is described in Table 4.1, where it is measured with respect to the number of abstracts represented and the number of sentences, mentions and tokens contained. Unique mentions refers to the number of unique mention texts. Disease concepts refers to the total number of unique disease concepts referenced in the corpus. Note there are approximately 1.75 unique mention texts per disease concept.

Several measurements were performed to gather descriptive statistics regarding the tokens, sentences and mentions in the corpus. Figure 4.1 shows the distribution of the number of tokens per mention. Figure 4.2 shows the distribution of mentions per sentence. Approximately 38% of the sentences

Figure 4.1: Number of tokens per mention in the Arizona Disease Corpus.

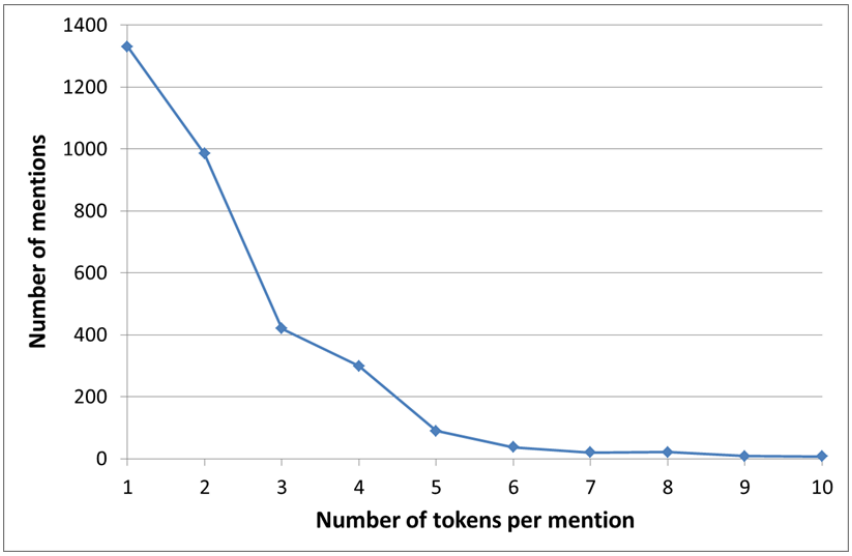


Figure 4.2: Number of sentences in the Arizona Disease Corpus containing a specific number of mentions.

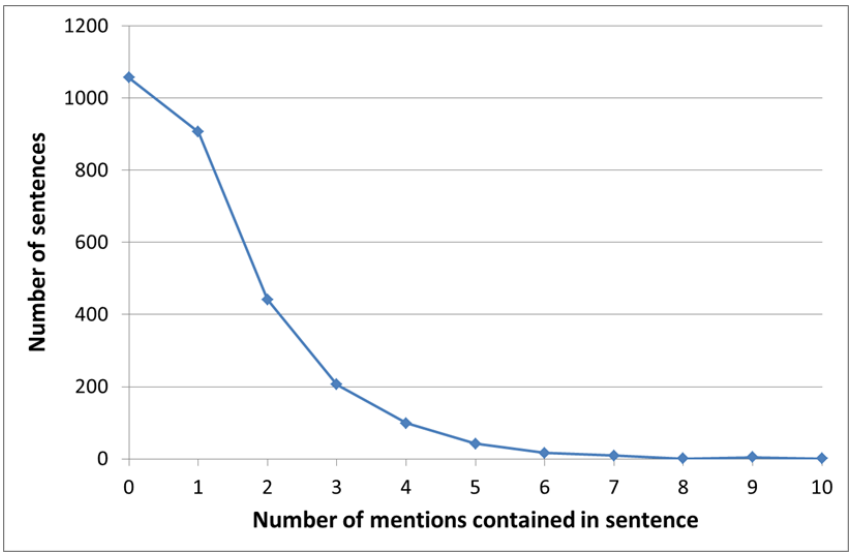
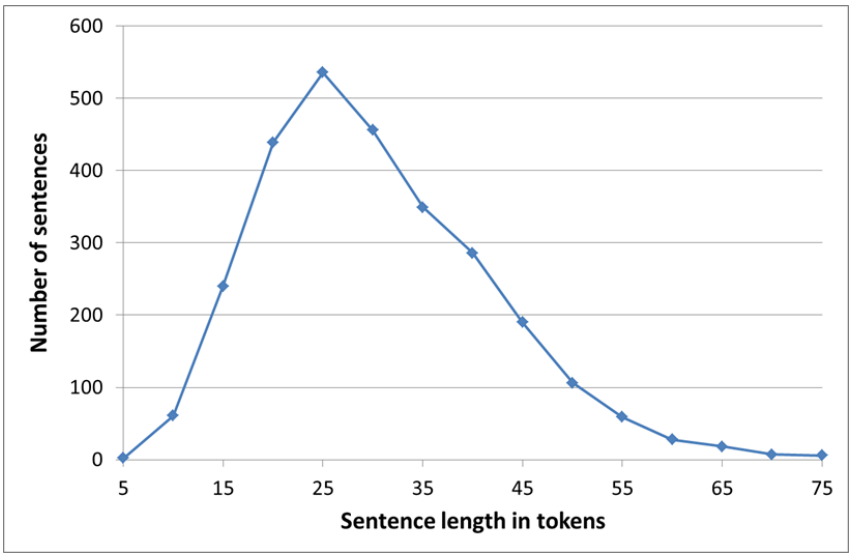


Figure 4.3: Distribution of sentence lengths in the Arizona Disease Corpus.

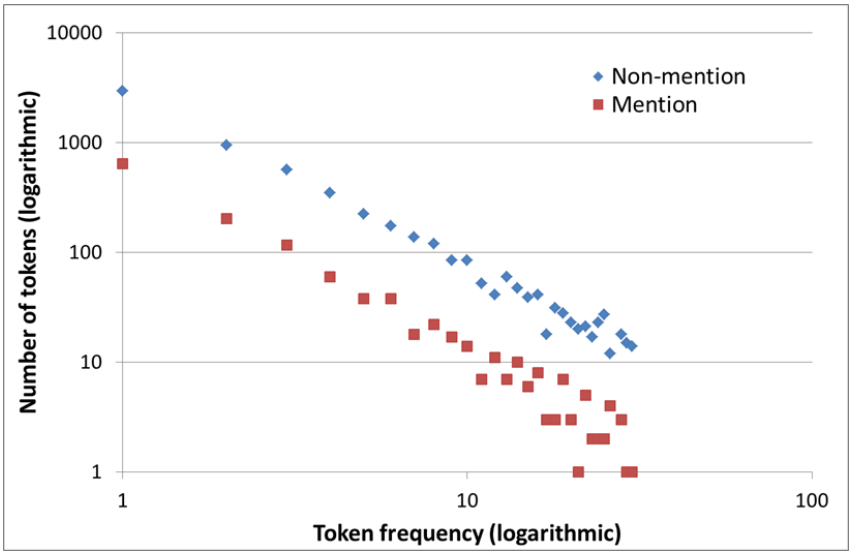


contain no disease mentions, which is useful since most sentences in arbitrary biomedical research text also do not contain disease mentions. Figure 4.3 describes the distribution of sentence lengths present in the corpus, which has median 25 and positive skew. Figure 4.4 shows the relationship between the number of tokens and the frequency with which they appear. This figure uses a log-log plot to highlight the adherence to Zipf's law [77, 128].

4.4.2 NER Results

BANNER and JNET were evaluated using 10-fold cross validation, splitting the corpus into 10 roughly equal parts, then training on 9 parts and then testing on the remaining 1, and repeating 10 times so that each part is used for testing once. The performance reported is then the average performance of all ten runs. Since sentences from the same abstract are much more likely to be similar than arbitrary sentences, the corpus is split so that all sentences from the same abstract were assigned to the same split. This ensures that

Figure 4.4: Distribution of tokens appearing in the Arizona Disease Corpus with the specified frequency.



there will never be a sentence in the training data and one in the test data which are from the same abstract. The dictionary method does not require training, and performance was therefore measured against one run of the entire corpus.

The exact matching criterion was used since it is the most strict matching criterion and therefore provides the most conservative estimate of performance. The precision, recall and f-measure were determined using the standard calculations. Table 4.2 summarizes the results of the named entity recognition study.

Dictionary lookup performed reasonably well, achieving a F-measure of 0.622. This is slightly better than the 0.592 F-measure found by Jimeno et al. (2008), which is probably attributable to the change in the evaluation corpus, as the current effort is over 4.5 times larger than the corpus employed by Jimeno et al.

System (Variant)	Precision	Recall	F-measure
Dictionary	0.627	0.617	0.622
BANNER (no dictionary)	0.785	0.699	0.740
BANNER (order 1)	0.795	0.744	0.768
JNET	0.824	0.727	0.772
BANNER (order 2)	0.809	0.751	0.779

Table 4.2: NER evaluation results for the dictionary method, three variants of BANNER, and JNET, using the exact match criterion and 10-fold cross validation.

Both of the machine learning systems significantly outperformed the dictionary method. The best performer was found to be BANNER with an F-measure of 0.779, with JNET following very closely at an F-measure of 0.772.

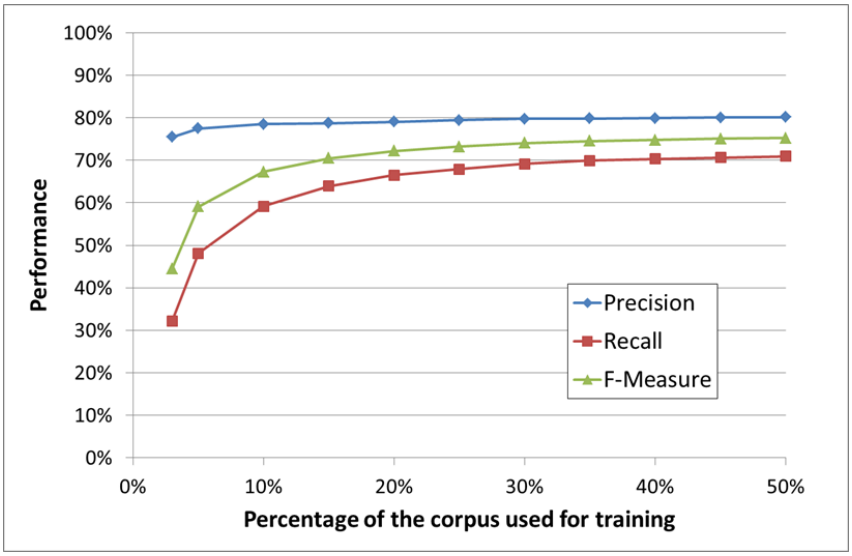
4.5 Discussion

An analysis of the number of disease concepts associated with each unique disease mention text shows that 91% of the texts are associated with exactly 1 disease concept, and 98% of them are associated with 2 or fewer disease concepts. This illustrates that ambiguity between disease concepts is low.

Polysemy is an issue, however, since many disease names, and especially their variations, map to the same disease concept. As a somewhat extreme example, familial adenomatous polyposis coli (C0032580), a condition where numerous polyps form in the colon that are initially benign but later turn malignant, is referred to by 13 different names in this corpus:

- 5 abbreviations: "AAPC," "APC," "FAP," "FAPC," and "FPC"
- 8 variations of the name "attenuated familial adenomatous polyposis coli" which include "polyposis" but leave out one or more other terms.

Figure 4.5: Ablation study using BANNER; the other 50% of the data was used for testing.



The effect of using differing amounts of training data was explored by using BANNER to perform an ablation study. The corpus was split into two equal parts, one for training and one for testing. 11 models were then trained using 5%, 10%, 20%, ...100% of the training data and then testing each on the entire test data. The model order was reduced to order 1, since it significantly reduces the training time required. These results are summarized in Figure 4.5. The ablation study demonstrated that while precision is hardly affected by the reduction in data, recall is significantly affected, and in fact drops precipitously at about 15% of the dataset and lower. The ablation study is also useful for extrapolating what performance could be achieved if the corpus were made larger. Analyzing the slope of both the precision and recall curves demonstrates that both flatten to nearly no slope as the percentage of the corpus used for training goes up. This suggests that, for BANNER at least, additional training data is not likely to markedly improve the performance achieved.

4.6 Error analysis

Output from the BANNER system was examined to investigate the errors that occurred and provide insight into the difficulties machine learning systems will experience in this domain. A subset of the corpus, consisting of 250 random sentences were reviewed in order to categorize the most common sources of error.

The difficulty which the system had in identifying acronyms is one notable source of error. The subset contained many instances of both false positives and false negatives due to this problem. Abbreviations in biomedical text are known to be a significant source of ambiguity [74]. Acronym disambiguation would improve accuracy to some degree, however this would need to be applied to each mention since the primary example of ambiguous acronyms in this corpus is the same text being used to refer to a disease and also to the associated gene from within the same abstract.

BANNER was also observed to not handle coordinations well, such as "Becker and Duchenne muscular dystrophy." Such phrases contain two or more mentions and are represented as such in the corpus. BANNER frequently only tags the part of the coordination containing the contiguous mention, which in the example would be "Duchenne muscular dystrophy." The AZDC contains 123 coordinations, with a total of 259 named entities. The technique of Buyko et al. for resolving coordinations was implemented and employed as a post-processing step prior to normalization candidate generation [15]. It was determined that while BANNER only tagged the complete coordination in 58 of the 123 sentences containing coordination ellipsis, the method still increased the number of correct concepts found by

the candidate generation step (which employed the Dice coefficient as a simple string similarity measurement based on sets of tokens) by approximately fourfold. While this method benefits normalization, the NER step should benefit from a set of features indicating the presence and boundaries of a coordination to improve the chance of capturing it in its entirety.

While named entities are typically a fixed phrase, it is often natural to refer to diseases by their effects rather than by name. Examples of this were apparent in the difficulties the system had in tagging mentions related to anatomical abnormalities and enzyme deficiencies. Anatomical abnormalities are often phrased descriptively, for example, "defect of the anterior midline scalp", "abnormality in ocular drainage structures" or "aberrantly developed trabecular meshwork." Mentions referring to enzyme deficiencies are similarly descriptive, for example, "lack of homogentisic acid oxidase." To correctly recognize these mentions, BANNER would benefit from a feature noting the location and boundaries of noun phrase chunks. These cases are relatively uncommon even so, and their recognition might also benefit therefore from enriching the dataset using, for example, the active-learning-like technique dynamic sentence selection [114].

BANNER is equipped with a post-processing module to detect and handle local abbreviations such as "adrenomyeloneuropathy (AMN)" [99]. However, this module expects the local abbreviation to be denoted by some form of parenthesis or bracket, which is a convention usually followed in biomedical text, though this corpus contains several exceptions. For example, in the sentence "A 40-year-old man with childhood-onset Tourette syndrome

(TS) developed Huntington disease HD.", BANNER should correctly handle "TS," but will likely not correctly handle "HD."

4.7 Conclusion

This chapter presents a new corpus of biomedical sentences, annotated both the location of disease mentions and the identity of the disease concepts. This corpus has been shown to enable the training of two machine-learning based named entity recognition systems, both of which, in turn, achieve higher performance than dictionary matching. Most of the difficulties with other biomedical entities have also been demonstrated to be relevant for diseases. Namely:

- There are a large number of names in use.
- Multiple names are used to refer to the same entity.
- Similar or identical names are used to refer to entities of differing semantic types.
- Complex syntactic structures, in this case coordinations, are used to refer to multiple related entities.

Intra-concept ambiguity, that is, the same name being used to refer to multiple disease concepts, is not a significant concern for disease recognition and identification, however.

Chapter 5

CASE STUDY: ADVERSE DRUG REACTIONS

It is estimated that approximately 2 million patients in the United States are affected each year by severe adverse drug reactions, resulting in roughly 100,000 fatalities. This makes adverse drug reactions the fourth leading cause of death in the U.S, following cancer and heart diseases [42]. It is estimated that $136 billion is spent annually on treating adverse drug reactions in the U.S., and other nations face similar difficulties [73, 117]. Unfortunately, the frequency of adverse drug reactions is often under-estimated due to a reliance on voluntary reporting [5, 117].

While severe adverse reactions have received significant attention, less attention has been directed to the indirect costs of more common adverse reactions such as nausea and dizziness, which may still be severe enough to motivate the patient to stop taking the drug. The literature shows, however, that non-compliance is a major cause of the apparent failure of drug treatments, and the resulting economic costs are estimated to be quite significant [58, 116]. Thus, detecting and characterizing adverse drug reactions of all levels of severity is critically important, particularly in an era where the demand for personalized health care is high.

An adverse drug reaction is generally defined as an unintended, harmful reaction suspected to be caused by a drug taken under normal conditions [71, 122]. This definition is sufficiently broad to include such conditions as allergic reactions, drug tolerance, addiction or aggravation of the original condition. A reaction is considered severe if it "results in death,

requires hospital admission or prolongation..., results in persistent or significant disability/incapacity, or is life-threatening," or if it causes a congenital abnormality [71].

The main sources of adverse drug reaction information are clinical trials and post-marketing surveillance instruments made available by the Food and Drug Administration (FDA), Centers for Disease Control and Prevention (CDC) in the United States, and similar governmental agencies worldwide. The purpose of a clinical trial, however, is only to determine whether a product is effective and to detect common serious adverse events. Clinical trials, by their nature and purpose, are focused on a limited number of participants selected by inclusion/exclusion criteria reflecting specific subject characteristics (demographic, medical condition and diagnosis, age). Thus, major uncertainties about the safety of the drug remain when the drug is made available to a wider population over longer periods of time, in patients with co-morbidities and in conjunction with other medications or when taken for off-label uses not previously evaluated.

Recently, the regulatory bodies of both the U.S. and the U.K. have begun programs for patient reporting of adverse drug reactions. Studies have shown that patient reporting is of similar quality to that of health professionals, and there is some evidence that patients are more likely to self-report adverse drug reactions when they believe the health professionals caring for them have not paid sufficient attention to an adverse reaction [9]. In general, however, the FDA advocates reporting only serious events through MedWatch.

Self-reported patient information captures a valuable perspective that might not be captured in a doctor's office, clinical trial, or even in the most

sophisticated surveillance software. For this reason, the International Society of Drug Bulletins asserted in 2005 that "patient reporting systems should periodically sample the scattered drug experiences patients reported on the internet."

Social networks focusing on health related topics have seen rapid growth in recent years. Users in an online community often share a wide variety of personal medical experiences. These interactions can take many forms, including blogs, microblogs and question/answer discussion forums. For many reasons, patients often share health experiences with each other rather than in a clinical research study or with their physician [28]. Such social networks bridge the geographical gap between people, allowing them to connect with patients who share similar conditions–something that might not be possible in the real world.

This chapter proposed and evaluated automatically extracting relationships between drugs and adverse reactions in user posts to health-related social network websites. This technique will provide valuable additional confirmation of suspected associations between drugs and adverse reactions. Moreover, it is possible this technique may eventually provide the ability to detect novel associations earlier than with current methods.

5.1 Related Work

In the work closest in purpose to this study, two reviewers manually analyzed 1,374 emails to the BBC and 862 messages on a discussion forum regarding a link between the drug paroxetine and several adverse reactions including withdrawal symptoms and suicide [80]. The authors concluded that the user

reports contained clear evidence of linkages that the voluntary reporting system then in place had not detected.

Not much work has been done to automatically extract adverse reactions from text, other than the SIDER side effect resource, which was created by mining drug insert literature [65]. There is, however, significant literature support for mining more general concepts, such as diseases. MetaMap is a primarily lexical system for mapping concepts in biomedical text to concepts in the UMLS Metathesaurus [4]. The ConText system categorizes findings in clinical records as being negated, hypothetical, or historical [51].

Most of the work on finding diseases concerns either biomedical text or clinical records. A notable exception is the BioCaster system, which detects infectious disease outbreaks by mining news reports posted to the web [22].

Health social networks have become a popular way for patients to share their health related experiences. A considerable amount of research has been devoted to this area [84], but most of this work has focused on the study of social interactions and quality evaluation instead of text mining. Automated information extraction from health social network websites remains largely unexplored.

5.2 Data Preparation

The DailyStrength[1] health-related social network was used as the source of user comments in this study. DailyStrength allows users to create profiles, maintain friends and join various disease-related support groups. It serves as a resource for patients to connect with others who have similar conditions,

[1] http://www.dailystrength.org

many of whom are friends solely online. As of 2007, DailyStrength had an average of 14,000 daily visitors, each spending 82 minutes on the site and viewing approximately 145 pages [24].

5.2.1 Data Acquisition

To efficiently gather user comments about specific drugs from the DailyStrength site, a highly parallelized automatic web crawler was implemented. All data was scraped from the raw HTML using regular expressions since the site has no open API. Users indicate a specific treatment when posting comments to DailyStrength, however treatments which are not drugs were filtered. For each user comment the user ID, disease name, drug name, and comment text were extracted. While more information about each user is available at the site (gender, age, self-declared location, and length of membership at the site), only the comment data were used. The DailyStrength Privacy Policy states that comments made by users will be publicly available. All data was gathered in accordance with the DailyStrength Terms of Service, and to respect fair use the data will not be made publicly available without permission from the site.

5.2.2 Preparing the Lexicon

To enable finding adverse reactions in the user comments, a lexicon was created by combining terms and concepts from four resources.

The UMLS Metathesaurus is a resource containing many individual biomedical vocabularies [87]. The subset used was limited to the COSTART vocabulary created by the U.S. Food and Drug Administration for

post-marketing surveillance of adverse drug reactions, which contains 3,787 concepts.

The SIDER side effect resource contains 888 drugs linked with 1,450 adverse reaction terms extracted from pharmaceutical insert literature [65]. The raw term found in the literature and the associated UMLS concept identifier (CUI) were used.

The Canada Drug Adverse Reaction Database, or MedEffect[2], contains associations between 10,192 drugs and 3,279 adverse reactions, which was used to create a list of adverse reaction terms. Many adverse reaction terms were found with very similar meanings, for example "appetite exaggerated," and "appetite increased," which were grouped together manually.

A small set of colloquial phrases were also included. These were collected manually from a subset of the DailyStrength comments and mapped to UMLS CUIs. This list is available[3], and includes the terms "throw up," meaning vomit, "gain pounds," meaning weight gain, and "zonked out," meaning somnolence.

All terms which are associated with the same UMLS concept identifier (CUI) as were considered to be synonymous and were grouped into a single concept. All concepts containing a term in common were also merged into a single unified concept. The lexicon contains 4,201 unified concepts, each containing between one and about 200 terms.

[2]http://www.hc-sc.gc.ca/dhp-mps/medeff/index-eng.php
[3]http://diego.asu.edu/downloads/adrs

Drug name (Brand name)	Primary Indications
carbamazepine (Tegretol)	epilepsy, trigeminal neuralgia
olanzapine (Zyprexa)	schizophrenia, bipolar disorder
trazodone (Oleptro)	depression
ziprasidone (Geodon)	schizophrenia
aspirin	pain, fever, reduce blood clotting
ciprofloxacin (Cipro)	bacterial infection

Table 5.1: List of drugs included in the subset for analysis and their primary indications.

5.3 Annotation

Comments relating to the following 4 drugs were annotated: carbamazepine, olanzapine, trazodone, and ziprasidone. These drugs were chosen because they are known to cause adverse reactions. The blood pressure medication clonidine was considered for inclusion, however it was eliminated from further consideration since a preliminary analysis demonstrated that many users confused it with the infertility drug clomifene. Comments for the drugs aspirin and ciprofloxacin were retained but not annotated; these comments are used during evaluation. These drugs are listed along with their primary indications in table 5.1. The data contains a total of 6,890 comment records. User comments were selected for annotation randomly and were independently annotated by two annotators.

Annotator 1 has a BS in biology, 10 years nursing experience in the behavioral unit of a long term care facility, and has dispensed all of the drugs annotated. Annotator 2 has a BS and an MS in neuroscience, and has work experience in data management for pharmaceutical-related clinical research and post-marketing drug surveillance.

Concept	Definition
Adverse effect	A reaction to the drug experienced by the patient, which the user considered negative
Beneficial effect	A reaction to the drug experienced by the patient, which the user considered positive
Indication	The condition for which the patient is taking the drug
Other	A disease or reaction related term not characterizable as one of the above

Table 5.2: The concepts annotated in this study and their definitions.

5.3.1 Concepts Annotated

Each comment was annotated for mentions of adverse effects, beneficial effects, indications and other terms, as defined in table 5.2. Each annotation included the span of the mention and the name of the concept found, using entries from the lexicon described in section 5.2.2. Each annotation also indicates whether it refers to an adverse effect, a beneficial effect, an indication or an other term, which shall be hereafter termed its characterization.

5.3.2 Annotation Practices

There are four aspects which require careful consideration when characterizing mentions. First, the stated concept may or may not be actually experienced by the patient; mentions of concepts not experienced by the patient were categorized as other. Second, the user may state that the concept is the reason for taking the drug. If so, the mention was categorized as an indication. Third, the concept may be an effect caused by the drug. In this case, the mention is categorized as either an adverse effect or a beneficial

effect based on whether the user considers the effect a positive one. This requires some judgment regarding what people normally view as positive – while sleepiness is normally an adverse effect, someone suffering from insomnia would consider it a beneficial effect, regardless of whether insomnia is the primary reason for taking the drug. Mentions of concepts which were experienced by the patient but neither an effect of the drug nor the reason for taking it were also categorized as other. Concepts were characterized as an adverse effect unless the context indicated otherwise.

Comments not containing a mention or that only indicated the presence of an adverse effect ("Gave me weird side effects") were discarded. If more than one mention occurred in a comment, then each mention was annotated separately.

Some comments clearly mentioned an adverse reaction, but the reaction itself was ambiguous. For example, in the comment "It did the job when I was really low. However, I BALLOONED on it," the annotator could infer "BALLOONED" to mean either weight gain or edema. A frequent example is colloquial terms such as "zombie," which could be interpreted as a physiological effect (e.g. fatigue) or a cognitive effect (e.g. mental dullness). In such cases, each mention was annotated by using both the context of the mention and annotator's knowledge of the effects of the drug.

Spans were annotated by choosing the minimum span of characters from the comment that would maintain the meaning of the term. Locating the mention boundaries was straightforward in many cases, even when descriptive words were in the middle of the term ("It works better than the other meds ive taken but I am gaining some weight"). However some comments were not as simple ("it works but the pounds are packing on").

Sample Comments	Annotations
hallucinations and weight gain	"hallucinations" - hallucinations: adverse effect; "weight gain" - weight gain: adverse effect
This has helped take the edge off of my constant sorrow. It has also perked up my appetite. I had lost a lot of weight and my doctor was concerned.	"constant sorrow" - depression: indication; "perked up my appetite" - appetite increased: beneficial effect; "lost a lot of weight" - weight loss: other
It worked well, but doctor didn't asked for the treatment to continue once my husband was doing well again.	none
ARGH! Got me nicely hypomanic for two weeks, then pooped out on me and just made me gain a half pound a day so I had to stop.	"hypomanic" - hypomania: beneficial effect; "pooped out" - tolerance: adverse effect; "gain a half a pound a day" - weight gain: adverse effect
Works to calm mania or depression but zonks me and scares me about the diabetes issues reported.	"mania" - mania: indication; "depression" - depression: indication; "zonks me" - somnolence: adverse effect; "diabetes" - diabetes: other
Works for my trigeminal neuralgia. Increasing to see if it helps stabalize mood. Fatigue!	"trigeminal neuralgia" - trigeminal neuralgia: indication; "stabalize mood" - emotional instability: indication; "Fatigue" - fatigue: adverse effect
Take for seizures and bipolar works well	"seizures" - seizures: indication; "bipolar" - bipolar disorder: indication
fatty patti!	"fatty" - weight gain: adverse effect

Table 5.3: An illustrative selection of uncorrected comments submitted to the DailyStrength health-related social networking website, and their associated annotations.

5.3.3 Corpus Description

A total of 3,600 comments were annotated, a sample of which can be seen in table 5.3. 450 comments were reserved for system development. The annotators found 1,260 adverse effects, 391 indications, 157 beneficial effects and 78 other, for a total of 1,886 annotations.

The agreement between annotators was measured by calculating both kappa (κ) [19] and inter-annotator agreement (IAA). For κ, agreement was considered to mean that the concept terms were in the same unified concept from the lexicon and the characterization of the mentions matched, since there is no standard method for calculating κ which includes the span. For IAA, an additional constraint was added that the annotation spans must overlap, since discussions of IAA typically include the span. Using these definitions, κ was calculated to be 85.6% and IAA to be 85.3%[4].

5.4 Text Mining

Since the drug name is specified by the user when the comment is submitted to DailyStrength, no extraction was necessary for drug names. To extract the adverse drug reactions from the user comments, a primarily lexical method was implemented, utilizing the lexicon discussed in section 5.2.2.

5.4.1 Methods Used

Each user comment was split into sentences using the Java sentence breaker, tokenized by splitting at whitespace and punctuation, and tagged for part-of-speech using the Hepple tagger [53]. Stop-words were removed from both user comments and lexical terms[5]. Tokens were stemmed using the Snowball implementation of the Porter2 stemmer[6].

Terms from the lexicon were found in the user comments by comparing a sliding window of tokens from the comment to each token in the lexical term. The size of the window is configurable and set to 5 for this study since that is the number of tokens in the longest term found by the

[4] κ>IAA here due to the different definitions of agreement.
[5] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words
[6] http://snowball.tartarus.org

annotators. Using a sliding window allows the tokens to be in different orders and for there to be irrelevant tokens between the relevant ones, as in weight gain and "gained a lot of weight."

Since user comments contain many spelling errors, the Jaro-Winkler measurement of string similarity was used to compare the individual tokens [121]. The similarity between the window of tokens in the user comment and the tokens in the lexical term was scored by pairing them as an assignment problem [14]. The similarities of the individual tokens was then summed and normalized the result by the number of tokens in the lexical term. This score is calculated for both the original tokens and the stemmed tokens in the window, and the final score is taken to be the higher of the two scores. The lexical term is considered to be present in a user comment if the final score is greater than a configurable threshold.

It was found that most mentions could be categorized by using the closest verb to the left of the mention, as in "taking for seizures." As this study focuses on adverse effects, a filtering method was implemented to remove indications, beneficial effects, and other mentions on a short list of verbs which indicate them. Verbs on this list include "helps," "works," and "prescribe" all of which generally denote indications. The complete list is available[7].

5.4.2 Text Mining Results

The system was first evaluated against the 3,150 annotated comments not reserved for system development. The evaluation was limited to adverse effects because the purpose is to find adverse drug reactions. This evaluation

[7] http://diego.asu.edu/downloads/adrs

used a strict definition of true positive, requiring the system to label the mention with a term from the same unified concept as the annotators. The results of this study are 78.3% precision and 69.9% recall, for an f-measure of 73.9%.

Since the purpose of this study is to determine if mining user comments is a valid way to find adverse reactions, the system was run on all available comments and compared the frequencies of adverse reactions found against their documented incidence. The frequency that each adverse effect was found in the user comments for each of the drugs studied in this experiment was counted. The most commonly found adverse reactions for each drug were then determined and compared against the most common documented adverse reactions for the drug. Since the four drugs chosen for annotation all act primarily on the central nervous system, aspirin and ciprofloxacin were added for this study. The results of this evaluation contain encouraging correlations that are summarized in table 5.4.

5.5 Discussion

The experiment comparing the documented incidence of adverse reactions to the frequency they are found contained some interesting correlations and differences. The adverse reaction found most frequently for all 6 of the drugs corresponded to a documented adverse reaction. There were also similarities in the less common reactions, such as diabetes with olanzapine and bleeding with aspirin. In addition, many of the adverse reactions found corresponded to documented, but less common, reactions to the drug. Examples of this included edema with olanzapine, nightmares with trazodone, weight gain with ziprasidone, tinnitus with aspirin, and yeast infection with ciprofloxacin.

Drug name (Brand name)	Documented Adverse Effects (Frequency)	Adverse Effects Found in User Comments (Frequency)
carbamazepine (Tegretol)	dizziness, somnolence or fatigue, unsteadiness, nausea, vomiting	somnolence or fatigue (12.3%), allergy (5.2%), weight gain (4.1%), rash (3.5%), depression (3.2%), dizziness (2.4%), tremor/spasm (1.7%), headache (1.7%), appetite increased (1.5%), nausea (1.5%)
olanzapine (Zyprexa)	weight gain (65%), alteration in lipids (40%), somnolence or fatigue (26%), increased cholesterol (22%), diabetes (2%)	weight gain (30.0%), somnolence or fatigue (15.9%), appetite increased (4.9%), depression (3.1%), tremor (2.7%), diabetes (2.6%), mania (2.3%), anxiety (1.4%), hallucination (0.7%), edema (0.6%)
trazodone (Oleptro)	somnolence or fatigue (46%), headache (33%), dry mouth (25%), dizziness (25%), nausea (21%)	somnolence or fatigue (48.2%), nightmares (4.6%), insomnia (2.7%), addiction (1.7%), headache (1.6%), depression (1.3%), hangover (1.2%), anxiety attack (1.2%), panic reaction (1.1%), dizziness (0.9%)
ziprasidone (Geodon)	somnolence or fatigue (14%), dyskinesia (14%), nausea (10%), constipation (9%), dizziness (8%)	somnolence or fatigue (20.3%), dyskinesia (6.0%), mania (3.7%), anxiety attack (3.5%), weight gain (3.2%), depression (2.4%), allergic reaction (1.9%), dizziness (1.2%), panic reaction (1.2%)
aspirin	nausea, vomiting, ulcers, bleeding, stomach pain or upset	ulcers (4.5%), sensitivity (3.8%), stroke (3.1%), bleeding time increased (2.8%), somnolence or fatigue (2.7%), malaise (2.1%), weakness (1.4%), numbness (1.4%), bleeding (1.0%), tinnitus (0.7%)
ciprofloxacin (Cipro)	diarrhea (2.3%), vomiting (2.0%), abdominal pain (1.7%), headache (1.2%), restlessness (1.1%)	abdominal pain (8.8%), malaise (4.4%), nausea (3.8%), allergy (3.1%), somnolence or fatigue (2.5%), dizziness (1.9%), weakness (1.6%), tolerance (1.5%), rash (1.3%), yeast infection (1.1%)

Table 5.4: List of drugs analyzed, with the 5 most common adverse effects, their frequency of incidence in adults taking the drug over the course of one year (if available) and the 10 most frequent adverse effects found in the the DailyStrength data using the automated system.

One interesting difference is the relative frequency of "hangover" in the comments for ziprasidone. Since the users were not likely referring to a literal hangover, they were probably referring to the fatigue, headache, dry mouth and nausea that accompany a hangover, all of which are documented adverse reactions to the drug.

Users frequently commented on weight gain and fatigue while ignoring other reactions such as increased cholesterol. While this may be because users are more conscious of issues they can directly observe, this hypothesis would not explain why other directly observable reactions such as nausea and constipation are not always reported. Determining the general trends in the differences between clinical and user reports is an important area for future work.

5.5.1 Error Analysis

An analysis was performed to determine the primary sources of error for the extraction system. 100 comments were randomly selected and determined the reason for the 24 false positives (FPs) and 29 false negatives (FNs) found.

The largest source of error (17% of FPs and 55% of FNs) was the use of novel adverse reaction phrases ("liver problem") and descriptions ("burn like a lobster"). This problem is due in part to idiomatic expressions, which may be handled by creating and using a specialist lexicon. This problem might also be partially relieved by the appropriate use of semantic analysis. However, this source of error is also caused by the users deliberately employing a high degree of linguistic creativity ("TURNED ME INTO THE SPAWN OF SATAN!!!") which may require deep background knowledge to correctly recognize.

The next largest source of error was poor approximate string matching (46% of FPs and 17% of FNs). While users frequently misspelled words, making lexical analysis difficult, the approximate string matching technique used also introduced many FPs. Spelling unfamiliar medical terminology is particularly difficult for users. Correcting this important source of error will require improved modeling of the spelling errors made by users.

Ambiguous terms accounted for 8% of the FPs and 7% of the FNs. While this is frequently a problem with colloquial phrases ("brain fog" could refer to mental dullness or somnolence), there are some terms which are ambiguous on their own ("numb" may refer to loss of sensation or emotional indifference). These errors can be corrected by improving the analysis of the context surrounding each mention.

Surprisingly, miscategorizations only accounted for 4% of the FPs. This small percentage seems to indicate that the simple filtering technique employed is reasonably effective. However this source of error can be seen more prominently in the frequency analysis, as seen in table 5.4. For example, one of the most frequent effects found in comments about trazodone was insomnia, which is one of its most common off-label uses. Other examples included depression with olanzapine, mania with ziprasidone, and stroke with aspirin. The remaining errors include one unrecognized term, "hungry," and a phrase which was spread across two sentences ("worked ... then stopped"). While this error is not common, the method may benefit from an extension to handle negation, since conditions not being experienced by the patient are always categorized as other.

5.5.2 Limitations

The present study has some limitations. The demographics of the users whose comments were mined were not analyzed, though it is likely that they are predominantly from North America and English-speaking. Future work should include expansion of the range of users and compare their demographics against clinical studies of adverse reactions. Also, the drugs annotated operate primarily on the central nervous system and therefore have different adverse reaction profiles than would other drugs with substantially different mechanisms. While the inclusion of aspirin and ciprofloxacin does provide some evidence these techniques are more generally applicable, the range of drugs studied should also be expanded in future work.

5.5.3 Opportunities for Further Study

In addition to the current classification for adverse reactions, there are additional dimensions along which each user comment could be studied. For example, many comments describe the degree of the adverse reaction, which can be straightforward ("extremely") or more creative ("like a pig"). Also, many users explicitly state whether they are still taking the drug, typically indicating whether their physician took them off or whether they took themselves off (non-compliance), and whether adverse reactions were the reason. User comments can also be categorized as medically non-descriptive ("I took one tablet and could'nt get out of bed for days and felt like I got hit by a truck"), somewhat medically descriptive ("My kidneys were not functioning properly"), or medically sound ("I ended up with severe leg swelling"). Comments also typically indicate whether the user is the patient

or a caretaker by being phrased in either the first person or third person narrative. Finally, users also frequently describe whether they thought the benefits of the drug outweighed the adverse effects. These additional dimensions represent a fertile area for further research.

5.6 Conclusion

In summary, user comments to health related social networks have been shown to contain extractable information relevant to pharmacovigilance. This approach should be evaluated for the ability to detect novel relationships between drugs and adverse reactions.

In addition to the improvements discussed in section 5.5, future work will increase the scale of the study (additional drugs, additional data sources, more user comments), improve the characterization of reactions using rule-based patterns, and evaluate the improved system with respect to all characterizations.

Chapter 6

MULTIVARIATE FEATURE SELECTION WITH FALSE DISCOVERY RATE CONTROL

Feature extraction creates many features of relatively low quality. This chapter now turns to improving the feature set quality by removing these poor features with feature selection. For a feature selection algorithm to be successful for biomedical NER, it should be able to handle extremely unbalanced distributions. The technique should also be able to handle high degrees of feature redundancy and control for multiple comparisons in a wide dataset. Moreover, the feature set created should be stable. In addition, since training conditional random fields requires significant computational resources, there is a strong preference for efficient techniques focus solely on filter-based feature selection in this chapter.

6.1 Related Work

Feature selection methods can be categorized into three groups: wrappers, filters and embedded methods [44]. Wrapper methods select features by alternating rounds of training a model and either adding or removing features. For example, recursive feature elimination trains a model using the entire feature set and then analyzes the model to select a set of features to remove because of low weight [46]. The second group, the filter methods, selects a set of features prior to training any model. These techniques typically rank the features according to a measure of relevance, and then select a subset by setting a threshold. Since these methods only train one model, they are considerably more computationally efficient than the

wrapper methods. The third type of feature selection methods are the embedded techniques. These methods include feature selection as part of the training process through techniques such as regularization.

The literature on feature selection outside of biomedical NER is vast, though most of the work is heuristic rather than theoretical. Only a few of the works most important for the present effort are presented, and the reader is referred to a reference for a more comprehensive discussion [45, 76].

A very recent work unifies many filter-based feature selection algorithms under a single framework [13]. The authors take a theoretical perspective of feature selection and use conditional likelihood maximization to derive the metric whose optimal value corresponds to the optimal feature set. The authors show that optimal feature selection must consider conditional relevance, that is, the relevance of a feature not already provided by other members of the feature set. This implies that removing redundant features need not always improve the classification performance. Moreover, the authors demonstrate that many successful feature selection heuristics are actually approximations of the full optimization problem derived by the authors. These include important algorithms such as fast correlation-based filtering (FCBF) [126], incremental association markov blanket (IAMB) [113], and mutual information maximization (MIM) [75]. The authors conclude that the joint mutual information (JMI) algorithm is the best approximation of the optimal objective function for accuracy and stability [82, 123].

6.1.1 Multiple comparisons

One of the most important concepts for feature selection is multiple comparisons, the fact that the probability of finding a statistically significant

association increases with the number of variables tested [60]. While multiple comparisons can be controlled using Bonferroni correction, this technique is conservative in the case of correlated or dependent variables and therefore not appropriate for every problem. Another technique, called the false discovery rate, has been introduced to allow controlling multiple comparison by limiting the percentage of features falsely considered relevant [8]. Assuming a large number of probes and a small number of relevant features, Guyon et. al. show that the ceiling for the false discovery rate can be estimated as follows:

$$FDR \leq \frac{n_{sp}}{n_p}\frac{n_c}{n_{sc}}$$

Where:

- n_{sp} is number of selected probes
- n_p is total number of probes
- n_c is total number of candidates
- n_{sc} is number of selected candidates

Probes are never added to the feature set and are therefore not used to build the final model. The false discovery rate can be used to stop feature selection when the false discovery rate reaches a specified point. The power of different feature selection algorithms for a given dataset can be compared by determining the number of features selected at different false discovery rates.

6.1.2 Feature Selection for NER

There have been relatively few evaluations of feature selection for biomedical NER. The first work evaluating feature sets for biomedical NER used

recursive feature elimination to reduce the feature set for a system using a support vector machine to classify tokens in a sliding window [49]. The authors demonstrate a 0.7% increase in the overall performance using the best feature set found using this method, and were also able to show that a performance only 2.3% lower than the maximum is achievable with less than 5% of the features.

A more recent study selected the most useful features for a biomedical NER problem by setting manual thresholds for the feature occurring and for the feature occurring inside a mention [83]. Features which fell below the thresholds were discarded and not used for training or evaluation, which would tend to increase both stability and generalization. The authors show an approximately 2% improvement in the f-measure using their technique.

The most recent work analyzes several feature selection methods on an NER system based on conditional random fields [62]. The authors evaluate the two most popular filtering methods, information gain and the χ^2 test, in addition to a wrapper technique, iterative feature pruning. The authors show that IG out-performs χ^2, and indicate a belief that this is due to the extreme class imbalance. Other research confirms that the χ^2 test is less accurate than information gain for very low counts [35]. The authors note that even the random baseline can be used to remove 30-40% of the original feature set without significantly impacting the achievable f-measure, implying a high degree of feature redundancy.

6.2 Methods

In a very recent work previously discussed in the survey of feature selection, the joint mutual information (JMI) feature selection score was shown to have

the best overall balance between considering feature relevance and feature interactions [13, 82, 123]. This technique is therefore applied to biomedical NER. As this technique already handles redundant features, extensions are proposed which allow high imbalance and false discovery rate analysis with probes.

JMI is a sequential search algorithm, adding one feature to the feature set at a time by selecting the feature with the highest JMI score. The JMI score is a function of the current feature set, and may increase or decrease as the feature set changes.

The JMI score for a feature not yet accepted can be calculated incrementally, so that the each round only requires a number of operations proportional to the number of features left to consider.

To determine when to stop adding new features, the feature selection is augmented with false discovery rate analysis using the probes technique, as described in the survey of feature selection. This is straightforward since JMI is a sequential search algorithm.

While NER features are binary, there is a significant imbalance between the number of true and false values for individual features. The probes should therefore not assume that the probability of a true and a false are approximately equal. Instead, the observed probability of a true value for each feature is calculated according to the feature set. To create the probes, the original feature set is assumed to be a set of binary random variables $x_1 \ldots x_n$ with observed probability of success $p_1 \ldots p_n$. A set of probes is created, $r_1 \ldots r_n$, each modeled by a Bernoulli distribution with probability of success chosen with replacement from $p_1 \ldots p_n$. Since the features are

highly imbalanced, the correct number of successes for each probe are guaranteed by instantiating a sparse vector of the same length as the number of tokens in the training set, ensure that it contains the correct number of successes, and then permute it.

6.3 Results

JMI with false discovery rate analysis was evaluated on the BioCreative 2 Gene Mention data and the NCBI Disease dataset. For the BioCreative 2 Gene Mention dataset, JMI selected 107 features prior to selecting any probes, for an estimated FDR of 0%. However the estimated FDR climbs to over 50% over the next 5 features accepted. For the NCBI Disease dataset, 57 features are selected prior to selecting any probes, again for an estimated FDR of 0%. In this case, however, only probes are selected after this point, so that to the estimated FDR climbs to over 50% with no more features selected.

A BANNER model was created using only the features selected by JMI with a FDR of 0%. Normally an FDR threshold of 0% would be too restrictive, but since the FDR climbs so quickly after that point it was not considered critical to set the threshold more carefully. The results can be seen in table 6.1.

6.4 Discussion

Using JMI as a feature selection method did not result in improved performance. For the BioCreative 2 Gene Mention corpus, precision dropped by 0.199 and recall dropped by more than 0.439. For the NCBI Disease corpus, precision dropped by 0.124 and recall dropped by more than 0.188. Given the very few features selected, a significant reduction in performance is

System (Variant)	Corpus	Precision	Recall	F-measure
BANNER, order 1	NCBI Disease	0.825	0.785	0.805
Using only features selected by JMI	NCBI Disease	0.701	0.617	0.656
BANNER, order 1	BioCreative 2 GM	0.860	0.834	0.847
Using only features selected by JMI	BioCreative 2 GM	0.635	0.408	0.496

Table 6.1: NER evaluation results for joint mutual information with FDR control.

not unexpected. It is also expected that fewer features would affect recall more than precision. Considering that that the full feature set constitutes hundreds of thousands of features, however, the performance achieved with three orders of magnitude fewer features is surprisingly high.

It is interesting that JMI selected only a few features before selecting mostly probes, causing the false discovery rate to climb quickly. The behavior of selecting probes almost exclusively after some point is somewhat interesting since it is so different from the behavior of univariate feature selection measures, whose probability of selecting a probe increases gradually as more features are accepted. This behavior is explained by the amount of new information remaining in the unselected features being less than the amount of new information apparently present in the probes. This is evidence, however, that the feature set created by the rich feature set approach has different properties than feature sets studied in other domains. It also suggests that the feature set returned by JMI may be too closely adapted to the training set, implying that the feature set itself is, in a sense, overtrained.

To test this hypothesis, 5 iterations of JMI over bootstrap samples were run, consisting of 50% of the BioCreative 2 Gene Mention dataset. JMI

was found to have repeated the behavior of selecting a number of features without selecting any probes, then selecting mostly probes. The result was an average of 83 features selected at 0% estimated FDR, with only 33 features present in all 5 models. This indicates a high degree of instability in the feature set chosen, and since the only difference between the data for the models was the training set, this is evidence that the feature set is too closely adapted to the training set.

6.5 Conclusion

In conclusion, joint mutual information was unable to select a higher performing feature set than simply selecting all features. However since linear-chain conditional random fields is based on a finite state machine, the feature set is different for each transition in the machine. The feature selection employed only allowed features to be available or unavailable for all transitions. It is possible that allowing features to be selected for each transition individually would enable a performance increase. It is also possible that the redundancy in the feature set is important, and that there are no features which can be removed while simultaneously supporting a performance improvement. Additional work in feature selection for named entity recognition may enable the performance improvements sought.

6.5.1 Future Work

Other recent work with false discovery rate analysis on high-dimensional data suggests that the technique stability selection may be useful. Stability selection creates a feature selection filter by combining many supervised classifiers into an ensemble [81]. Each model in the ensemble is trained using $\mathcal{L}_1$ regularization, causing the weight for many features to be pushed down to

exactly zero, which is equivalent to removing these features from the feature set.

The authors show that the probability that a feature has a nonzero weight in the ensemble models is both highly correlated with the feature being relevant and is highly stable. In addition, the technique is not sensitively dependent on the choice of the $\mathcal{L}_1$ regularization parameter, and the probability threshold can be varied to control the false discovery rate. The resulting feature set is then used to create a final classification model.

The analysis allowing FDR calculations assumes that the features are independent and identically distributed, which is not accurate, and the results of the experiment with JMI suggests that this assumption may be problematic. In addition, when the feature set contains several useful features that are highly correlated, the algorithm will usually select none of them. The reason is that $\mathcal{L}_1$ regularization arbitrarily selects only one feature from a set of duplicate features, resulting in none of them being selected frequently enough individually. This can be solved using the elastic net – a linear combination of $\mathcal{L}_1$ and $\mathcal{L}_2$ regularization that results in a sparse feature set that is significantly more stable [129].

Chapter 7

INCORPORATING LEXICONS THROUGH LANGUAGE MODELING

This chapter describes a novel method for characterizing the content of token sequences by adapting methods from terminology extraction and language modeling. The purpose is to enable the creation of a novel technique for named entity recognition, intended to be used as a feature. The technique proposed utilizes language modeling as a way to describe how common specific sequences are in a corpus of interest, and uses the difference between the probability of the sequence appearing in two corpora to characterize the likelihood that this sequence appeared in one of them.

The purpose of this technique is to allow more efficient use of domain resources such as lexicons as features for named entity recognition. While many entities of interest lack lexicons and other domain resources, such resources do exist for many of the most important entities, including genes, proteins, diseases, known genomic variations, species, and many more. Given the significant investment required to create such domain resources, it would be useful to take advantage of them when they are available.

7.1 Related work

These methods are strongly inspired by work in terminology extraction, which is the problem of locating terms, or multi-token sequences, which refer to entities of interest in a particular domain. In terminology extraction the output is a list of terms, analogous to a list of unique mentions in NER. However in terminology extraction, the locations that each term was found and the semantic type is not considered relevant. Terms are, however,

frequently given a score representing how strongly the sequence appears to be a term in the text of interest.

One classic paper in terminology extraction uses the relative frequency ratio to locate terms of interest by comparing the relative frequency of a term in a foreground corpus with the relative frequency of that term in a background corpus [26]. The foreground corpus is taken to be a large amount of text known to be relevant to the domain of interest, while the background text is taken to be a large amount of unlabeled text from a wider domain. The relative frequency ratio is defined as follows:

$$R_w(p||q) = \frac{p(\mathbf{w})}{q(\mathbf{w})}$$

Where:

- $p(\mathbf{w})$ is the probability of the phrase $\mathbf{w}$ appearing in the forground corpus
- $q(\mathbf{w})$ is the probability of the phrase $\mathbf{w}$ appearing in the background corpus

The authors were able to demonstrate that the relative frequency ratio was effective in locating bigrams, expressions of length two, that are important for the chosen domain.

The relative frequency ratio has been criticized as being too sensitive to the frequency of the term in the background corpus, giving especially high weight to terms which appear infrequently or not at all, regardless of their frequency in the foreground corpus. More recent work extends the idea of comparing the distributions of phrases between two corpora by using the pointwise Kullback-Leibler (KL) divergence, which the authors define as

describing informativeness [111]. The pointwise KL divergence is defined as the amount contributed by a single phrase **w** to the KL divergence, and is calculated as follows:

$$\delta_w(p||q) = p(\mathbf{w})log\frac{p(\mathbf{w})}{q(\mathbf{w})}$$

Where:

- $p(\mathbf{w})$ is the probability of the phrase **w** appearing in the forground corpus
- $q(\mathbf{w})$ is the probability of the phrase **w** appearing in the background corpus

The pointwise KL divergence is a measurement of the information lost when the phrase is considered to have come from the background corpus, modeled by q(x), when it actually came from the foreground corpus, modeled by p(x). The authors show that this metric has properties useful for key phrase extraction, including prioritizing tokens whose frequency is high in both the foreground and background.

Another popular technique, the C-value / NC-value method, incorporates both a measurement of termhood and an analysis of context to determine which sequences refer to terms [40]. The termhood of a phrase is taken to be the frequency of occurrence, minus the frequency as a substring within other terms. The final designation as a term also incorporates an analysis of the context, where potential terms are used to find context words that are likely triggers and these are then used to reinforce the final values for termhood for each term. The authors note that this statistical technique requires reevaluation by domain experts and filtering. This technique has

been employed in a biomedical context, and the idea of also taking advantage of the context where terms appear will also be employed by us – using a different method – in the next chapter.

7.1.1 Survey of language modeling

Language modeling estimates the probability of observing a specified sequence of tokens, usually a sentence. Language modeling has found uses in many well-established natural language processing tasks, particularly machine translation [106], though it is also finding use in new research areas, such as weakly-supervised techniques [57] and active learning [33]. A strong advantage of language models is that they do not require labeled training data. Rather, the only requirement is for a large quantity of the text to be modeled, a condition satisfied in biomedical text through MEDLINE abstracts and the PubMed open access subset of full-text articles.

One of the most common techniques for creating a language model is n-grams. An n-gram model predicts the next token in a sequence, given n-1 previous tokens [77]. For example, n=3 in a trigram model, and the model uses the previous 2 tokens to predict the following token. These probabilities can be chained to calculate the probability of an arbitrarily long sequence. This is often done in log space so as to prevent underflow.

A fundamental problem in language modeling is that there are valid sequences which will not be observed in the training data; that is, the training data is sparse. A straightforward application of maximum likelihood estimation will result in a zero probability for all sequences not seen previously. This is typically corrected through smoothing, which increases the probability of unseen items slightly by reducing the probability of

sequences observed. A simple illustrative example is Laplace smoothing, also called add-one smoothing, because all counts are considered to be one higher than the count observed. This results in the probability of a sequence never being zero, but also results in most of the probability mass in the distribution being reserved for unseen tokens, which may not be desirable [77]. Many other techniques have been studied, including linear interpolation, Good-Turing discounting, and Katz Backoff [17].

Building a wide-coverage n-gram model requires the storage of billions of individual counts. To reduce the space requirements, language models frequently employ a probabilistic data structure called a Bloom filter [10]. Bloom filters allow a space efficient way to store counts such that any errors are one-way. That is, the structure may return a count higher than it should be (false positive), but never a count that is lower (false negative). Recent work employing Bloom filters in language models use a log-scale version as a drop-in replacement for the filter with exact counts, resulting in significant additional space savings [106, 107].

Other techniques for language modeling use maximum entropy modeling [95], or model the entire sequence as a single instance [96]. Recent work in language modeling explores many possible improvements, including incorporating syntactic information [16], exploiting knowledge of the topic of the document [108], and using a vector space of word similarity to reduce the data sparsity problem [7]. Several studies performing empirical evaluations demonstrate consistently strong results with a trigram model using interpolated Kneser-Ney smoothing [17, 43], and ignoring words that occur only once, also known as hapax legomena [120].

7.2 Methods

Machine-learning based NER systems have frequently attempted to use features incorporating additional domain knowledge. Lists of entity names, frequently called dictionaries, are a domain resource that is not available for all entity types of interest, due to the expense of creating and maintaining the resource. When a list of entity names is available, however, there is a strong desire to make use of the resource. Some early attempts to incorporate dictionaries into biomedical NER systems resulted in reduced performance, however [100]. Previous work showed a small improvement in the performance after adding a name list consisting of single tokens filtered to only contain names highly indicative of an entity name [50].

This chapter describes a method to describe the degree to which a token sequence resembles an entity name. The degree to which a sequence resembles an entity name is determined by taking the difference between two language models, resulting in a real-valued feature. To do this a language modeling approach designed for keyphrase extraction is adapted that models the informativeness of a term as the loss between two language models: one language model for the domain or application and one that is general. The main idea of this technique is that language modeling allows us to create a model to predict the likelihood of an arbitrary string. In this experiment, however, two models are employed: a foreground model representing the sorts of sequences which should be located and a background model that reflects the sequences which should be ignored. Intuitively, the difference between the probability estimates of these two models reflects the likelihood that the given sequence comes from either the foreground or background model.

System (Variant)	Corpus	Precision	Recall	F-measure
BANNER, order 1	NCBI Disease	0.820	0.784	0.802
With the feature	NCBI Disease	0.810	0.780	0.794
BANNER, order 1	BioCreative 2 GM	0.860	0.834	0.847
With the feature	BioCreative 2 GM	0.869	0.830	0.849

Table 7.1: NER evaluation results for the method of characterizing sequences with language modeling, across two corpora.

In this application, a domain lexicon is employed as the foreground text and MEDLINE, a large unlabeled corpus of biomedical abstracts, as the background text. The likelihood of a token sequence referring to a entity is defined as the pointwise Kullback-Leibler divergence between the probability of the sequence according to the language model for the lexicon and the probability according to the language model for the large unlabeled biomedical corpus. The MEDIC disease lexicon is used for diseases [27], while EntrezGene is used for genes and proteins [86]. Both the unsmoothed language models and Laplace-smoothed language models are used. Also, both the relative frequency ratio and the pointwise KL divergence are considered.

7.3 Results

This technique was evaluated using the BioCreative 2 Gene Mention data and the NCBI Disease Corpus, as described in Table 7.1. Because the MEDIC lexicon is relatively small, the creation of a unigram model is described. Higher order models were attempted, however they were not successful.

7.4 Discussion

Given that modern language modeling techniques employ corpora with millions or even billions of tokens, it was considered likely that domain lexicons would be too sparse to learn an effective language model, since they

contain thousands of tokens. This was not the case, however. Domain lexicons are intended to be relatively complete for the entity described, implying that much of the vocabulary needed to recognize entities of that type will be present, even if many specific term variants are not. Moreover, should it become necessary to deal with the sparseness in ways other than smoothing, it is possible to use models of approximate word meaning rather than probability of word appearance [7].

A more pressing concern is the fact that the meaning of the token being frequent in the lexicon is not the same as the meaning of a token being frequent in natural language text. With diseases, the technique is strong for locating words that are frequently used to describe diseases, but not as strong at locating the head word itself. For example, in the phrase "autosomal recessive disease," both "autosomal" and "recessive" are given a relatively high score, while "disease" is given a relatively low score. This occurs because even though the word "disease" appears frequently in the foreground corpus, the lexicon, it also appears frequently in the background corpus, MEDLINE, decreasing the difference between their distributions. Many other disease names have similar issues. An excellent example is the word "tumor," which often refers to a disease but is also frequently used in expressions describing other related concepts, such as "tumor suppressor." The method was not observed to prefer either heads or descriptive words for genes and proteins.

This technique has several significant advantages. First, while it does require the use of a lexicon, these are often available. More importantly, this technique is unsupervised and therefore does not require training data. Second, while state-of-the-art systems employing language modeling require significant resources due to the use of very large corpora or long n-grams

[107], the use of language models requires several orders of magnitude fewer resources. This technique is therefore very lightweight. Moreover, the advanced techniques described in the literature to enable high performance from the much larger language models can also be employed here. Third, because there has been significant research into techniques to improve language models, there is a significant amount of existing literature that can be explored in an attempt to improve this technique.

7.5 Conclusion

A novel method for named entity recognition has been proposed by repurposing techniques used in language modeling and term recognition. While this approach was not able to demonstrate a performance improvement, variations of this technique may deserve further study.

Future named entity recognition studies employing this technique would likely benefit from using multiple instances to model many lexicons simultaneously. This would be useful because it would allow the machine learning model to determine whether a the resemblance to another entity type increases or decreases the likelihood that it refers to the type of interest. This approach may be useful for studies with only a small amount of training data. In an active learning approach this technique would provide fast initial improvements to the system by allowing it to concentrate primarily on the differences between the way the mentions are represented in the lexicon and the way they are used in text, rather than needing to learn both from the beginning. Finally, it would be interesting to evaluate expanding this technique to other comparisons besides a lexicon as a foreground corpus and a domain text is a background corpus. For example, it would be expected

that biomedical entities should appear in a domain corpus, such as MEDLINE, more frequently than in a nonspecific English corpus like newswire.

Chapter 8

CHARACTERIZING SEQUENCES WITH DISTRIBUTIONAL SEMANTICS

The rich feature set approach is characterized by employing a large number of features, each of which contains a relatively small amount of information. This chapter describes a method to improve the feature set by introducing a new technique for named entity recognition based on distributional semantics and employing it as a feature. A form of the distributional semantics hypothesis is employed, namely that it is possible to approximate the meaning of a word by observing the contexts in which it appears. Such techniques are advantageous for NER since they can partially compensate for the finite size of the training data, and have proved popular in recent work [61, 115]. Unlike previous work in distributional semantics, however, the method proposed models the meaning of token sequences as a unit rather than modeling the meaning of individual tokens. This is important since entity names exhibit meanings that are non-compositional, meaning that they cannot be accurately predicted from the sum of the meaning of their constituent tokens [77]. MEDLINE abstracts are used as the unlabeled text, since biomedical abstracts often contain statements defining the entities in the discourse.

8.1 Related work

Distributional semantics approaches have been used successfully to improve performance of named entity recognition systems by improving the representation of existing features or introducing new ones [61, 115].

Distributional semantics is based on the distributional hypothesis, which states that the meaning of a word can be inferred from the contexts in which it appears [52, 54]. The idea was famously summarized by Firth as "you shall know a word by the company it keeps" [38]. Distributional semantics is therefore the creation of quantitative models of word meaning through analyzing the contexts where words are found [21].

Distributional semantics techniques can be broadly characterized as either probabilistic or geometric. Probabilistic models of distributional semantics treat texts as a mixture of several topics so that the probability that a word appears can be modeled as the combination of the probability of each topic. Geometric models represent words as vectors in a high-dimensional space created as a representation of the contexts where words appear. Another important property of distributional semantic models is the type of relationships represented [29]. A syntagmatic relationship is the type of relationship between words that tend to co-occur, such as between a gene name and the disease caused by a variant of the gene, or perhaps a protein and its subcellular localization. A paradigmatic relationship is between words that can substitute for the other without modifying the syntactic structure of the sentence, such as between the names of different genes.

A significant advantage of distributional semantics is that it does not require expensive labeled data, only a relatively large amount of unlabeled text. There are many words which do not appear frequently enough in the training data to gain a sense of their meaning, but do appear frequently enough in a large unlabeled text to approximate their meaning with distributional semantics. Existing approaches in distributional semantics

model the meaning of single tokens, even though the names of many biomedical entities span multiple tokens.

8.2 Methods

In preliminary experiments, it was determined that the context surrounding mentions of both genes and diseases exhibits distinctive variations from the remainder of the text. Dunning's likelihood ratio is used to determine which tokens appear significantly more often in the context either to the left or to the right of an entity mention more frequently than would be expected by chance. Many thousands of tokens were found to appear more frequently than would be expected by chance when all mentions of a specified type were considered as a single group. However, when each mention individually was considered individually, it was rare to find any tokens which appear significantly more frequently than would be expected by chance. This is primarily because the frequency of each individual mention is so low.

While these results are not unexpected, it would be beneficial to find some tokens appear more frequently than would be expected by chance around many mentions, thereby giving high confidence that these tokens are highly indicative of entities of that type. It was therefore determined instead to select the tokens that appear most frequently when mentions are considered as a single group. It was decided to follow prior work and employ K nearest neighbors.

The method for using unlabeled data to create features for biomedical NER proceeds as follows. For each mention in the labeled data, all instances of the same sequence of tokens in the unlabeled data were located. For each instance of the same sequence of tokens, the context to the left and the right

of the mention were extracted, and the frequency of each token found tallied. Once all mentions were processed, the counts of each token were compared with the count expected by chance, based on the number of times the token appears in the corpus.

To summarize the method, labeled token sequences from the training data are represented as vectors created from the context surrounding other instances of that sequence in a large amount of unlabeled data. Rather than model the context on both sides simultaneously, one model is created for the right side and another model for the left. The likelihood that any given unlabeled token sequence refers to a mention can then be determined by converting it into right and left context vectors, applying the K nearest neighbor algorithm to classify each, and then combining the result.

The initial step is preprocessing the text. Because this method is based on locating other instances of term surface forms, the main considerations for ensuring adequate performance are to reduce ambiguity and variation. The unlabeled text is broken into sentences using the Java sentence breaker, which was adapted to not break sentences within parentheses. Punctuation and stopwords are both dropped. Tokens are stemmed and numbers are normalized to the single digit zero, to improve handling of variation in surface forms. All abbreviations in the training and test data are resolved using the Schwartz and Hearst algorithm, to reduce the ambiguity of the terms [99].

Some tokens appearing in the context surrounding a potential mention may be significantly more useful for discriminating between mentions and non-mentions. Feature selection is therefore employed to limit the tokens used for inference to a relatively small set of the most useful. Dunning's

likelihood ratio test for collocations was adapted to find words appearing more frequently than chance surrounding token sequences [77]. Two feature sets are created, one for the context to the left of the potential mention, and one for the context to the right. All instances of the token sequences labeled as mentions in the training data are then located in the unlabeled data set . Next, the number of times each token appears in the context of a specified size surrounding the other instances is counted. These frequencies are compared with the frequency that each token appears in the corpus as a whole using Dunning's likelihood ratio. This statistic allows us to determine the number of times the token appears more frequently than chance, and is known to be more accurate at low counts than Pearson's Chi-squared test [91]. A threshold is set manually and select all tokens that are more likely to appear in the context surrounding a mention than the threshold.

Two separate context vector models are then created, one for the context to the left of the sequence and the other representing the context to the right. Both models use an order-independent bag of words to represent individual tokens found within a window of specified size. The vectors are therefore very high dimensional (over one million), but also very sparse, typically containing less than a few hundred nonzero dimensions. The TF-IDF representation is used, where the TF is defined as the number of contexts where the token appears, and IDF is the log of the number of times the token appears in the entire unlabeled corpus. This representation reflects both the importance of the context token for the token sequence and also the relative importance of the context token in the corpus of the whole. All vectors are normalized to unit length to compensate for the wide variation in the frequency of each token sequence.

Each model is then loaded with labeled vectors using sequences from the training data. One vector is created for each mention in the training data. Vectors are also created for each token sequence up to the specified maximum mention length that is not part of a mention. While many approaches to terminology extraction use a white-list approach to determining which sequences are worth considering, this experiement used a black-list approach with only two rules, ignoring only sequences that are very unlikely to represent a mention. The first rule stipulates that the sequence may not begin or end with a word that is a member of a closed syntactic class in English, such as determiners. The second rule states that the sequence may not contain a comma. This rule was adopted empirically, to increase precision, since commas are valid in mentions involving coordination. Because token sequences may appear many times in the training data, the vectors were labeled with both the number of times that the token sequence represented appears as a mention in the training data, and also the number of times that the token sequence appears as a non-mention. Note that tokens in the training data that are not part of a mention will be used to create many vectors from overlapping n-grams, but each mention is only represented once, as a complete unit. Note also there are no vectors spanning the boundary between mention and non-mention text.

Arbitrary unlabeled token sequences can then be classified by using the large unlabeled corpus to convert it to a context vector and using k nearest neighbors with the labeled vectors from the training data. Standard cosine similarity is used as the similarity metric. Since each training data vector is annotated with the number of times its origin token sequence is labeled as a mention or as non-mention text, several methods were

considered to convert these counts into an overall summary value between zero and one. The conversion that was found to best differentiate between the sequences representing mentions and non-mentions is the average of the probability that each vector would be labeled as a mention, weighted by similarity. This score is considered to be the likelihood that the n-gram represents a mention. The likelihood from both the left and right models is obtained and converted to a joint likelihood by multiplying the two.

Given a sentence, the most likely segmentation for the sentence must be found. This can be determined using a greedy approach. The likelihood that each subsequence in the sentence refers to a mention according to the left and right models is determined. the sequences considered are limited in a manner similar to the training data, namely, sequences that either begin or end with tokens that are a member of a closed syntactic class, or that contain a comma, are not considered. The sequences are sorted in order of decreasing likelihood and consider each for selection in that order. Sequences that overlap with a previously selected sequence but do not contain it are removed. The result is a set of likely mentions, with some mentions (e.g. "male breast cancer") containing other likely mentions ("breast cancer"). All mentions above a predefined threshold are accepted and returned to the next processing unit. Alternatively, the highest score for each token in the sentence may be calculated and passed as features to a named entity recognizer such as BANNER.

8.3 Results

This technique was evaluated using the BioCreative 2 Gene Mention data and the NCBI Disease Corpus as training and test data. MEDLINE was used as

Genes and Proteins	
Left	Right
express	cell
cd0	express
activ	active
serum	gene
alpha	receptor
beta	protein
a	alpha
c	inhibitor
anti	level
human	ml
degre	respect
protein	mrna

Diseases	
Left	Right
breast	cell
human	patient
patient	mellitu
prostat	line
primari	viru
malign	necrosi
lung	a
cell	plyori
colorect	associ
doubl	cancer
type	suppressor
acut	factor

Table 8.1: List of stemmed tokens selected from those most strongly associated with appearing to the left or to the right of either genes or diseases.

the unlabeled corpus. The context window was set to 3 tokens. The feature selection only considered tokens that are 100 times more likely to appear in the context surrounding a disease mention than to appear in general.

These settings resulted in 9,754 tokens in the left context feature set and 10,252 tokens for the right context feature set. The stemmed tokens associated most strongly with appearing in the context of gene names and disease names, to both the left and the right, can be seen in Table 8.1. To reduce the search space, the maximum mention length was set to 6 tokens, resulting in the extraction of 276,473 labeled sequences from the training set. The number of nearest neighbors considered for the K nearest neighbors algorithm was set to 20.

The results of incorporating the method of characterizing sequences with distributional semantics as a feature into the BANNER named entity recognizer are described in Table 8.2. Both precision and recall increase by

System (Variant)	Corpus	Precision	Recall	F-measure
BANNER, order 1	NCBI Disease	0.820	0.784	0.802
With the feature	NCBI Disease	0.831	0.795	0.813
BANNER, order 1	BioCreative 2 GM	0.860	0.834	0.847
With the feature	BioCreative 2 GM	0.878	0.847	0.862

Table 8.2: NER evaluation results for the method of characterizing sequences with distributional semantics, across two corpora.

over 1.0% across both corpora. F-measure increases by 1.1% for diseases and 1.5% for genes. We performed an additional round of evaluation, limiting the base feature set to the features selected by joint mutual information with false discovery rate control. The results of this experiment are described in Table 8.3. We note that the addition of the feature characterizing sequences with distributional semantics results in significant performance increases, though the performance is not yet approaching the level of the full feature set.

8.4 Discussion

The first observation is that this method is adept at finding the head words of a mention, but not quite as strong at finding the mention boundaries. This is likely because the head words appear more frequently in MEDLINE and in more discriminative contexts. The score assigned to unlabeled context vectors according to the algorithm appears to decay exponentially rather than linearly. This is also not unexpected, given the high number of dimensions and the fact that the quality of the context vectors goes down as the frequency of the n-gram used to create it drops exponentially according to Zipf's law.

System (Variant)	Corpus	Precision	Recall	F-measure
BANNER, using only the features selected by JMI	NCBI Disease	0.701	0.617	0.656
BANNER, using the features selected by JMI and the distributional semantics feature	NCBI Disease	0.741	0.647	0.691
BANNER, using only features selected by JMI	BioCreative 2 GM	0.635	0.408	0.496
BANNER, using only features selected by JMI and the distributional semantics feature	BioCreative 2 GM	0.778	0.616	0.687

Table 8.3: NER evaluation results for the method of characterizing sequences with distributional semantics, across two corpora, using only the features selected by joint mutual information with FDR control.

8.4.1 Limitations

The most significant limitation of this technique is the processing time required to create and classify the context vectors. Tt required nearly one hour to create the context vectors for 100 PubMed abstracts. The amount of time required for classifying the vectors, however, varies significantly with the number of vectors created and also the number of tokens used as features in the context. Reducing the n-grams considered to lengths no longer than 6 and that do not begin or end with words from closed classes in English significantly reduced the number of n-grams considered, and was necessary to make the problem tractable. This technique may be employed in the future in conjunction with a technique that analyzes the content of the potential

mentions to determine which sequences are likely to contain entities of the specified type, thereby reducing the number of vectors that must be compared. This study will motivate additional research to find more computationally efficient ways to achieve similar results.

Another issue with this technique is that there is no ability to deemphasize a head found as part of a multi-word phrase that should be ignored. For example, the word "tumor" from "tumor suppressor" will initially receive the same score as "tumor" from "pancreatic tumor". This is partially corrected when the greedy method for filtering the sequences extends to "tumor" from "pancreatic tumor" the score from "pancreatic tumor," which should be much higher than "tumor" alone.

Next, this technique assumes that the meaning of sequences is fairly fixed throughout the literature. That is, the same sequence of tokens has the same meaning everywhere. This assumption is unfortunately not the case either in general or with entity names, however it is a closer approximation with entity names.

The length limitation is a limitation in two ways. First, the probability of a sequence being a sufficiently fixed phrase that it appears multiple times in the unlabeled corpus goes down significantly as the length increases. Second, the number of sequences to consider from the training data goes up exponentially in the length of the sequence. It is possible that this problem can be worked around by considering sequences of all lengths in a sentence but only using the first tokens on the left or last tokens on the right if a sequence is over a predetermined length. Evaluation of this potential improvement is left as future work.

8.5 Conclusion

Distributional semantics has been demonstrated to be a viable technique for improving named entity recognition when the compositionality of the mentions is considered. Moreover, the strong performance of this technique suggests that it may be viable, with additional work, alongside the conditional random fields approach rather than only as an additional feature.

To accomplish the goal of making this useful in the future, the primary limitation is the significant computational resources required for classifying each context vector. There are methods, however, which may increase the performance of this technique considerably. For example, while support vector machines and logistic regression require some time to train, they perform inference much faster than K nearest neighbor. They would therefore be strongly preferable, provided that a method such as cross validation is used to ensure that the sequence classification model trained on their output does not overfit. Moreover, if K nearest neighbors is retained, it would be useful to employ a dimensionality reduction technique such as latent semantic indexing or random indexing, both of which create much smaller vectors [21, 97].

The results of the right and left models would be ideal for incorporating as features in a semi-Markov conditional random fields system [98]. In semi-Markov conditional random fields, features may be defined against sequences, in addition to individual tokens. This would allow the output of the context vector classifier to be used directly as a feature in a semi-Markov conditional random fields model, bypassing the greedy sequence analysis algorithm employed.

This method may be extended to find indicative token sequences in the context rather than single tokens. It would be expected to find, for example, that "transcription factor" and "tumor suppressor" are both highly correlated with a protein mention. Augmenting the selection of highly indicative features with false discovery rate analysis using the probes technique, as described in the survey of feature selection, may be useful determining when to stop adding new features.

Chapter 9

CONCLUSION

The proposed techniques and evaluation have demonstrated the size of the performance improvements possible through improving the feature set through adding new features based on counts from unlabeled text and feature selection.

Biomedical named entity recognition using conditional random fields and the rich feature set approach has been very successful. These techniques have allowed named entity recognition performance to approach human level, and has reduced the effort required for creating a named entity recognition system for a new domain considerably. While this success is a welcome development, research in named entity recognition has concentrated primarily on this approach, neglecting other forms of information extraction and natural language processing. In a sense, this dissertation is an attempt to re-envision named entity recognition not as a nearly-solved problem employing highly specialized techniques, but rather demonstrate how advances in other areas of NLP can be profitably adapted and employed in named entity recognition.

9.1 Conclusions

In conclusion, it is very difficult to improve on conditional random fields with the rich feature set approach. While some of the features generated this way may not be stable, the feature set generated is of sufficient quality that there are as yet no known methods to improve on the set with feature selection. The techniques for modeling the content of potential mentions using lexicons

and language modeling was marginally successful, and may be worth revisiting going forward. However the technique for using distributional semantics to model the meaning of a potential mention was significantly successful, and with performance improvements, may become a standard part of named entity recognition systems such as BANNER.

9.2 Summary of Advances

Methods for improving the performance of biomedical named entity recognition (NER) systems have been considered. The fact that biomedical NER is an important real world problem with significant remaining difficulties has been noted. The state-of-the-art techniques for biomedical NER have been surveyed, including the rich feature set approach, and argued that this approach is subject to diminishing returns. Improving the feature set employed with two complementary approaches has therefore been proposed. In the first approach, two new feature templates for modeling the semantics of a token sequence, one template modeling context with distributional semantics, and the other modeling content with language modeling. In addition, two novel variations on feature selection techniques, with improvements for the specific needs of biomedical NER. These techniques consider overlapping and interacting features, can be used in environments with extremely high skew, and limit the number of irrelevant features admitted via false discovery rate control.

REFERENCES

[1] Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richart Tobin, and Xinglong Wang. Assisted curation: Does text mining really help? In Proceedings of the Pacific Symposium on Biocomputing, pages 556–567, 2008.

[2] Ethem Alpaydin. Introduction to Machine Learning. MIT Press, 2004.

[3] Cecilia N. Arighi, Zhiyong Lu, Martin Krallinger, Kevin Bretonnel Cohen, W. John Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy H. Wu. Overview of the BioCreative III Workshop. BMC Bioinformatics, 12 Suppl 8(Suppl 8):S1, 2011.

[4] Alan R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, pages 17–21, 2001.

[5] David W. Bates, R. Scott Evans, Harvey Murff, Peter D. Stetson, Lisa Pizziferri, and George Hripcsak. Detecting adverse events using information technology. Journal of the American Medical Informatics Association, 10(2):115–128, 2003.

[6] William A. Baumgartner Jr., Kevin Bretonnel Cohen, Lynne M. Fox, George Acquaah-Mensah, and Lawrence Hunter. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics, 23:i41–i48, 2007.

[7] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137–1155, 2003.

[8] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 57(1):289–300, 1995.

[9] A. Blenkinsopp, M. Wang, P. Wilkie, and P. A. Routledge. Patient reporting of suspected adverse drug reactions: a review of published literature and international experience. British Journal of Clinical Pharmacology, 63(2):148–156, 2007.

[10] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. Communications of the ACM, 13(7):422–426, 1970.

[11] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In Proceedings of the 6th Workshop on Very Large Corpora, New Brunswick, New Jersey, 1998.

[12] Philip E. Bourne and Johanna McEntyre. Biocurators: Contributors to the world of science. PLoS Computational Biology, 2(10):e142, 2006.

[13] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. Journal of Machine Learning Research, 13:27–66, 2012.

[14] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. Assignment Problems. Society for Industrial and Applied Mathematics, 2009.

[15] Ekaterina Buyko, Katrin Tomanek, and Udo Hahn. Resolution of coordination ellipses in biological named entities using conditional random fields. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 163–171, 2007.

[16] C. Chelba. Structured language modeling. Computer Speech & Language, 14(4):283–332, 2000.

[17] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4):359–393, 1999.

[18] Hong-Woo Chun, Jin-Dong Kim, Jun'ichi Tsujii, Naoki Nagata, Rie Shiba, Teruyoshi Hishiki, and Yoshimasa Tsuruoka. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. In Proceedings of the Pacific Symposium on Biocomputing, pages 4–15, 2006.

[19] Jacob Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1):37–46, 1960.

[20] Kevin Bretonnel Cohen, Tom Christiansen, William A. Baumgartner Jr., Karin Verspoor, and Lawrence E. Hunter. Fast and simple semantic class assignment for biomedical text. In Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, pages 38–45, 2011.

[21] Trevor Cohen and Dominic Widdows. Empirical distributional semantics: Methods and biomedical applications. Journal of Biomedical Informatics, 42(2):390–405, 2009.

[22] Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics, 24(24):2940–2941, 2008.

[23] Nigel Collier and Chikashi Nobata. Extracting the names of genes and gene products with a hidden Markov model. In Proceedings of the International Conference on Computational Linguistics, pages 201–207, 2000.

[24] comScore Media Metrix Canada. Key Measures Report - Health. Technical report, comScore Media Metrix Canada, Toronto, Ontario, Canada, 2007.

[25] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology, pages 77–86, 1999.

[26] Fred J. Damerau. Generating and evaluating domain-oriented terms from texts. Information Processing & Management, 29(4):433–447, 1993.

[27] Allan Peter Davis, Thomas C. Wiegers, Michael C. Rosenstein, and Carolyn J. Mattingly. MEDIC: A practical disease vocabulary used at the Comparative Toxicogenomics Database. Database, 2012:bar065, 2012.

[28] K. P. Davison, J. W. Pennebaker, and S. S. Dickerson. Who talks? The social psychology of illness support groups. The American Psychologist, 55(2):205–217, 2000.

[29] Ferdinand de Saussure. Cours de Linguistique Générale. 1922.

[30] Dina Demner-Fushman, Wendy W. Chapman, and Clement J. McDonald. What can natural language processing do for clinical decision support? Journal of Biomedical Informatics, 42(5):760–72, 2009.

[31] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7):1895–1923, 1998.

[32] Shipra Dingare, Malvina Nissim, Jenny Rose Finkel, Christopher D. Manning, and Claire Grover. A system for identifying named entities in biomedical text: How results from two evaluations reflect on both

the system and the evaluations. Comparative and Functional Genomics, 6(1-2):77–85, 2005.

[33] Dimitriy Dligach and Martha Palmer. Good seed makes a good crop: Accelerating active learning using language modeling. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 6–10, 2011.

[34] Rezarta Islamaj Dogan and Zhiyong Lu. An improved corpus of disease mentions in PubMed citations. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pages 91–99, 2012.

[35] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61–74, 1993.

[36] Richárd Farkas. The strength of co-authorship in gene name disambiguation. BMC Bioinformatics, 9(1):69, 2008.

[37] Jenny Rose Finkel, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. Exploring the boundaries: Gene and protein identification in biomedical text. BMC Bioinformatics, 6(Suppl 1):S5, 2005.

[38] John Rupert Firth. Papers in Linguistics 1934–1951. Oxford University Press, London, 1957.

[39] Radu Florian, John F. Pitrelli, Salim Roukos, and Imed Zitouni. Improving mention detection robustness to noisy input. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 335–345, 2010.

[40] Katerina T. Frantzi, Sophia Ananiadou, and Jun'ichi Tsujii. The C-value / NC-value method of automatic recognition for multi-word terms. Research and Advanced Technology for Digital Libraries, pages 585–604, 1998.

[41] Kristofer Franzén, Gunnar Eriksson, Fredrik Olsson, Lars Asker, Per Liden, and Joakim Coster. Protein names and how to find them. International Journal of Medical Informatics, 67(1-3):49–61, 2002.

[42] K. M. Giacomini, R. M. Krauss, D. M. Roden, M. Eichelbaum, M. R. Hayden, and Y. Nakamura. When good drugs go bad. Nature, 446(7139):975–977, 2007.

[43] Joshua T. Goodman. A bit of progress in language modeling, extended version. Technical report, Microsoft Research, 2001.

[44] Isabelle Guyon and Andre Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.

[45] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A. Zadeh, editors. Feature Extraction: Foundations and Applications. Springer Verlag, 2006.

[46] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir N. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3):389–422, 2002.

[47] U. Hahn, E. Buyko, R. Landefeld, M. Muhlhausen, M. Poprat, K. Tomanek, and J. Wermter. An overview of JCoRe, the JULIE Lab UIMA Component Repository. In Proceedings of the Workshop Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, pages 1–7, 2008.

[48] Udo Hahn, Katrin Tomanek, Elena Beisswanger, and Erik Faessler. A proposal for a configurable silver standard. In Proceedings of the Fourth Linguistic Annotation Workshop, pages 235–242, Uppsala, Sweden, 2010.

[49] Jörg Hakenberg, Steffen Bickel, Conrad Plake, Ulf Brefeld, Hagen Zahn, Lukas Faulstich, Ulf Leser, and Tobias Scheffer. Systematic feature evaluation for gene name recognition. BMC Bioinformatics, 6(Suppl 1):S9, 2005.

[50] Jörg Hakenberg, Conrad Plake, Robert Leaman, Michael Schroeder, and Graciela Gonzalez. Inter-species normalization of gene mentions with GNAT. Bioinformatics, 24(16):i126, 2008.

[51] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics, 42(5):839–851, 2009.

[52] Zellig S. Harris. Distributional Structure. Word, 10:146–162, 1954.

[53] Mark Hepple. Independence and commitment: assumptions for rapid training and execution of rule-based POS taggers. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pages 277–278, 2000.

[54] Donald Hindle. Noun classification from predicate-argument structures. In Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, pages 268–275, 1990.

[55] Lynette Hirschman, Gully A. P. C. Burns, Martin Krallinger, Cecilia Arighi, K. Bretonnel Cohen, Alfonso Valencia, Cathy H. Wu, Andrew Chatr-Aryamontri, Karen G. Dowell, Eva Huala, Analia Lourenco, Robert Nash, Anne-Lise Veuthey, Thomas Wiegers, and Andrew G. Winter. Text mining for the biocuration workflow. Database, 2012, 2012.

[56] Doug Howe and Seung Yon. The future of biocuration. Nature, 455(September):47–50, 2008.

[57] Fei Huang, Alexander Yates, Arun Ahuja, and Doug Downey. Language models as representations for weakly-supervised NLP tasks. In Proceedings of the 15th Conference on Computational Natural Language Learning, pages 125–134, 2011.

[58] Dyfrig A. Hughes, Adrian Bagust, Alan Haycox, and Tom Walley. The impact of non-compliance on the cost-effectiveness of pharmaceuticals: a review of the literature. Health Economics, 10(7):601–615, 2001.

[59] International Society Of Drug Bulletins. Berlin Declaration on Pharmacovigilance. Technical Report January, International Society Of Drug Bulletins, Berlin, Germany, 2005.

[60] David D. Jensen and Paul R. Cohen. Multiple comparisons in induction algorithms. Machine Learning, 38:309–338, 2000.

[61] Siddhartha Jonnalagadda, Robert Leaman, Trevor Cohen, and Graciela Gonzalez. A distributional semantics approach to simultaneous recognition of multiple classes of named entities. In Alexander Gelbukh, editor, 11th International Conference on Intelligent Text Processing and Computational Linguistics, pages 224–235, 2010.

[62] Roman Klinger and Cristoph M. Friedrich. Feature subset selection in conditional random fields for named entity recognition. In Galia Angelova, Kalina Bontcheva, Nicolai Nikolov, Nicolas Nicolov, and Ruslan Mitkov, editors, Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, pages 185–191, Borovets, Bulgaria, 2009.

[63] Roman Klinger and Katrin Tomanek. Classical probabilistic models and conditional random fields. Technical report, Technische Universität Dortmund, Dortmund, Germany, 2007.

[64] Corinna Kolárik, Roman Klinger, Christoph M. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. Chemical names: Terminological resources and corpora annotation. In Workshop on Building and Evaluating Resources for Biomedical Text Mining, pages 51–58, 2008.

[65] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. Molecular Systems Biology, 6:343–348, 2010.

[66] Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. Integrated annotation for biomedical information extraction. In Proceedings of Biolink 2004: Linking Biological Literature, Ontologies and Databases, pages 61–68, 2004.

[67] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning, pages 282–289, 2001.

[68] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 504–513, 2010.

[69] Robert Leaman and Graciela Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition. In Proceedings of the Pacific Symposium on Biocomputing, pages 652–663, 2008.

[70] Robert Leaman, Christopher Miller, and Graciela Gonzalez. Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark. In 2009 Symposium on Languages in Biology and Medicine, 2009.

[71] Anne Lee, editor. Adverse Drug Reactions. Pharmaceutical Press, second edition, 2006.

[72] Florian Leitner, Scott A. Mardis, Martin Krallinger, Gianni Cesareni, Lynette A. Hirschman, and Alfonso Valencia. An overview of BioCreative II.5. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 7(3):385–399, 2010.

[73] Roberto Leone, Laura Sottosanti, Maria Luisa Iorio, Carmela Santuccio, Anita Conforti, Vilma Sabatini, Ugo Moretti, and Mauro Venegoni. Drug-related deaths: An analysis of the Italian Spontaneous Reporting Database. Drug Safety, 31(8):703–713, 2008.

[74] Ulf Leser and Jörg Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. Briefings in Bioinformatics, 6(4):357, 2005.

[75] David D. Lewis. Feature selection and feature extraction for text categorization. pages 212–217, 1992.

[76] Huan Liu and Hiroshi Motoda, editors. Computational Methods of Feature Selection. Chapman & Hall/CRC, 2007.

[77] Christopher D. Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[78] Andrew McCallum. MALLET: A Machine Learning for Language Toolkit, 2002.

[79] Ryan T. McDonald, R. Scott Winters, Mark Mandel, Yang Jin, Peter S. White, and Fernando Pereira. An entity tagger for recognizing acquired genomic variations in cancer literature. Bioinformatics, 20(17):3249–3251, 2004.

[80] Charles Medawar, Andrew Herxheimer, Andrew Bell, and Shelley Jofre. Paroxetine, Panorama and user reporting of ADRs: Consumer intelligence matters in clinical practice and post-marketing drug surveillance. International Journal of Risk & Safety in Medicine, 15(3/4):161–169, 2002.

[81] Nicolai Meinshausen and Peter Bühlmann. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473, 2010.

[82] Patrick E. Meyer and Gianluca Bontempi. On the use of variable complementarity for feature selection in cancer classification. In Applications of Evolutionary Computing, pages 91–102. Springer Verlag, 2006.

[83] Pabitra Mitra, Sudeshna Sarkar, and Sujan Kumar Saha. Feature selection techniques for maximum entropy based biomedical named entity recognition. Journal of Biomedical Informatics, 42(5):905–911, 2009.

[84] S. T. Moturu, H. Liu, and W. G. Johnson. Trust evaluation in health information on the World Wide Web. In Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 1525–1528, 2008.

[85] Skatje Myers, Robert Leaman, and Graciela Gonzalez. Improving named entity recognition with deep parsing. Technical report, Computing Research Association, 2010.

[86] National Library of Medicine. EntrezGene.

[87] National Library of Medicine. UMLS Knowledge Sources, 2008.

[88] Aurélie Névéol, W. John Wilbur, Won Kim, and Zhiyong Lu. Exploring two biomedical text genres for disease recognition. In Proceedings of the Workshop on Biomedical Natural Language Processing, pages 144–152, 2009.

[89] Andrew Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the 21st International Conference on Machine Learning, pages 78–85, 2004.

[90] Mark Palatucci and Andrew Carlson. On the chance accuracies of large collections of classifiers. In Proceedings of the 25th International Conference on Machine Learning, pages 744–751. ACM, 2008.

[91] Ted Pedersen. Fishing for exactness. In Proceedings of the South Central SAS Users Group Conference, 1996.

[92] Sandeep Pokkunuri, Ellen Riloff, Marina Rey, and Eduard Hovy. The role of information extraction in the design of a document triage application for biocuration. In Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, pages 46–55, 2011.

[93] Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch, and Antonio Jimeno. Text processing through Web services: calling Whatizit. Bioinformatics, 24(2):296–298, 2008.

[94] Barbara Rosario and Marti A. Hearst. Classifying semantic relations in bioscience texts. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 430, 2004.

[95] Ronald Rosenfeld. A maximum entropy approach to adaptive statistical language modelling. Computer Speech & Language, 10(3):187–228, 1996.

[96] Ronald Rosenfeld, Stanley F. Chen, and Xiaojin Zhu. Whole-sentence exponential language models: A vehicle for linguistic-statistical integration. Computer Speech & Language, 15(1):55–73, 2001.

[97] Magnus Sahlgren. An introduction to random indexing. In Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, 2005.

[98] Sunita Sarawagi and William W. Cohen. Semi-markov conditional random fields for information extraction. In Advances in Neural Information Processing, volume 17, pages 1185–1192, 2004.

[99] Ariel S. Schwartz and Marti A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In Proceedings of the Pacific Symposium on Biocomputing, volume 8, pages 451–462, 2003.

[100] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pages 104–107, 2004.

[101] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srevro, and Karthik Sridharan. Learnability, stability and uniform convergence. Journal of Machine Learning Research, 11:2635–2670, 2010.

[102] Christian Siefkes. A Comparison of tagging strategies for statistical information extraction. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pages 149–152, 2006.

[103] D. D. Sleator and D. Temperley. Parsing English with a link grammar. Technical report, Carnegie Mellon University, 1991.

[104] Larry Smith, Lorraine K. Tanabe, Rie J. Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner Jr., Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña López, Jacinto Mata, and W. John Wilbur. Overview of BioCreative II gene mention recognition. Genome Biology, 9(Suppl 2):S2, 2008.

[105] Larry H. Smith and W. John Wilbur. The value of parsing as feature generation for gene mention recognition. Journal of Biomedical Informatics, 42(5):895–904, 2009.

[106] David Talbot and Miles Osborne. Randomised language modelling for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pages 512–519, 2007.

[107] David Talbot and Miles Osborne. Smoothed Bloom filter language models: Tera-scale LMs on the cheap. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 468–476, 2007.

[108] Ming Tan, Wenli Zhou, Lei Zheng, and Shaojun Wang. A scalable distributed syntactic, semantic and lexical language model. Computational Linguistics, (November), 2011.

[109] Lorraine Tanabe and W. John Wilbur. A priority model for named entities. In Proceedings of HLT-NAACL BioNLP Workshop, pages 33–40, 2006.

[110] Lorraine Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, and W. John Wilbur. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics, 6 Suppl 1:S3, 2005.

[111] Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pages 33–40, 2003.

[112] Richard Tzong-Han Tsai, Cheng-Lung Sung, Hong-Jie Dai, Chuan Hung, Ting-Yi Sung, and Wen-Lian Hsu. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC Bioinformatics, 14:1–14, 2006.

[113] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. Algorithms for large scale markov blanket discovery. In Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference, pages 376–381, 2002.

[114] Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. Accelerating the annotation of sparse named entities by dynamic sentence selection. In Proceedings of the Workshop on Biomedical Natural Language Processing, pages 30–37, 2008.

[115] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 384–394, 2010.

[116] John Urquhart. Pharmacoeconomic consequences of variable patient compliance with prescribed drug regimens. PharmacoEconomics, 15(3):217–228, 1999.

[117] Cornelis S. van Der Hooft, Miriam C. J. M. Sturkenboom, Kees van Grootheest, Herre J. Kingma, and Bruno H. Ch. Stricker. Adverse drug reaction-related hospitalisations: A nationwide study in the Netherlands. Drug Safety, 29(2):161–168, 2006.

[118] Andreas Vlachos. Evaluating and combining biomedical named entity recognition systems. In Proceedings of the Workshop on Biomedical Natural Language Processing, pages 199–206, 2007.

[119] Andreas Vlachos. Tackling the BioCreative2 gene mention task with conditional random fields and syntactic parsing. Proceedings of the Second BioCreative Challenge Evaluation Workshop, pages 85–87, 2007.

[120] Marc Weeber, Rein Vos, and R. Harald Baayen. Extracting the lowest-frequency words: Pitfalls and possibilities. Computational Linguistics, 26(3):301–317, 2000.

[121] William E. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, US Bureau of the Census, Washington, DC, 1999.

[122] World Health Organization. International drug monitoring: The role of the hospital. Technical report, World Health Organization, Geneva, Switzerland, 1966.

[123] Howard Hua Yang and John Moody. Data visualization and feature selection: New algorithms for nongaussian data. In Advances in Neural Information Processing Systems, 1999.

[124] Yiming Yang. Sampling strategies and learning efficiency in text categorization. In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access, pages 88–95, 1996.

[125] Alexander Yeh, Alexander A. Morgan, Marc Colosimo, and Lynette Hirschman. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics, 6 Suppl 1:S2, 2005.

[126] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning, 2003.

[127] Guodong Zhou, Dan Shen, Jie Zhang, Jian Su, and Soonheng Tan. Recognition of protein / gene names from text using an ensemble of classifiers. BMC Bioinformatics, 7:1–7, 2005.

[128] George Kingsley Zipf. Human behavior and the principle of least effort. Addison-Wesley Press, Oxford, England, 1949.

[129] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.